2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

The frailty model versus the Andersen-Gill model for the prediction of recurrent events

Promotor :
dr. Philippe HALDERMANS

Promotor :
Dr. YVONNE VERGOUWE
Prof. EWOUT STEYERBERG

Fitsum Megersa Baye

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten in twee landen: de Universiteit Hasselt en Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Maastricht University

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

The frailty model versus the Andersen-Gill model for the prediction of recurrent events

Promotor :
dr. Philippe HALDERMANS

Promotor :
Dr. YVONNE VERGOUWE
Prof. EWOUT STEYERBERG

Fitsum Megersa Baye

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Prediction of Recurrent Events in Bladder Cancer Patients: A Comparison of Marginal Models and Conditional Frailty Model

By:
Baye Fistum Megersa

Internal supervisor: dr. Haldermans Philippe
External Supervisors: dr. Vergouwe Yvonne
Prof. Steyerberg Ewout

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Biostatistics of Hasselt University

September 12, 2011

Certification

I declare that this thesis was written by me under the guidance and counsel of my supervisors.

.....	Date.....
Baye Fitsum Megersa	Student

We certify that this is the true thesis report written by Baye Fitsum Megersa under our supervision and we thus permit its presentation for assessment.

.....	Date.....
dr. Haldermans Philippe	Internal Supervisor

.....	Date.....
dr. Vergouwe Yvonne	External Supervisor

.....	Date.....
Prof. dr. Steyerberg Ewout	External Supervisor

Acknowledgements

This work would not have been accomplished without the assistance and support of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, I am heartily thankful to my supervisors, dr. Vergouwe Yvonne and dr. Haldermans Philippe, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. I appreciate the department of Public health at Rotterdam Erasmus Medical center, for letting me to work with them and to use their office during this study. I would like to thank all my professors at UHasselt and Censtat staff members, for giving me all the skills and valuable set of knowledge which will definitely shape my future career. My special thanks goes to VLIR (Vlaamse Inter Universitaire Raad) scholarship, for making my two years study possible by providing funds.

I want to thank my parents, my father, Megersa Baye, my mother, Seniat Desta, my sisters, my brothers and all family members, without whom I would never have been able to achieve this much. Most importantly, I wish to thank my loving and supportive fiance, Hambisa Assefa, whose love and support are with me all the time.

Above all, I thank God for listening to all my prayers and helping me to pass through the difficult times. All my successes and accomplishments were made possible by Him.

Last but not least, thank you all my best friends for your unconditional love and supports. "Mariam Tsewa Mahiber" members, you have been my other families in Belgium, thank you so much. May God bless you all.

September 12, 2011

Abstract

Recurrent events data have been increasingly important in clinical studies where individuals experience an event more than once. Cancer patients commonly experience recurrent of tumors. There is a growing interest in the analysis of recurrent events data in order to assess the relationship of relevant predictors to the rate in which events are occurring. Several statistical methods have been proposed to analyze such type of data. Often analyzing the time-to-first event survival analysis ignores possible subsequent events and results bias. The aim of this study is to compare different marginal and random effect survival methods to analyze time to multiple recurrences of bladder cancer and enable for future prediction. A total of 615 bladder cancer patients were followed up between the years 1974-2011 in different hospitals in Rotterdam and 1280 observations were recorded. Zero to a maximum of 15 recurrences were observed during this follow up time.

In this report, mainly four marginal and random-effect models, extended from the proportional Cox model were fitted. The Andersen-Gill, marginal and conditional models were compared from the marginal model family. The main differences among them lies on the time scale and risk set formulations. Counting process, 'gap' and total time scales were used to fit the models. These models only correct the standard errors for their assumption of independence between events from the same patient. Further, a gamma shared frailty model to account for intra-cluster correlation or unobserved heterogeneity is fitted. The predictive accuracy and discriminative ability of the models were compared using concordance probability (C-index). Results indicate that, the Andersen-Gill and the conditional models provided approximately similar and unbiased results compared to the marginal model. The frailty model has the highest estimated concordance probability than the other models suggesting that it has the highest predictive accuracy and discriminatory power. Gender, tumor multiplicity, and number of previous recurrence events were identified as important prognostic factors.

In summary, the Andersen-Gill and the frailty models were proposed as the most appropriate methods to model the recurrent event data. Thus, one can use either of the models for predicting absolute risks of bladder cancer depending on his/her interest.

Key words: *Andersen-Gill, Bladder cancer, Concordance probability, Counting process, Cox model, Frailty model, Marginal model, Recurrent event*

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

A great deal of research is currently underway to develop risk prediction models to accurately estimate the absolute risk of a specific disease. Absolute risk is the probability that an individual with given risk factors will develop the disease over a defined period of time. Prediction models combine patient characteristics to predict medical outcomes and to support decision making. Developing such statistical models will help clinicians identify individuals at higher risk of specific disease, allowing for earlier or more frequent counseling of behavioral changes to decrease risk [1]. These types of models are also useful for designing future intervention in individuals at high risk of the disease in the general population. In clinical studies, both single and repeated events are commonly used to develop risk prediction models for a particular disease.

Repeated event processes, where individual subjects or units under consideration experience the same or different types of events more than once over time are called recurrent events. In many scientific investigations, the outcome variable of interest is a recurrent event. Recurrent event data are ubiquitous across a great range of diverse fields such as medicine, public health, insurance, social science, economics, manufacturing and reliability [2]. Many epidemiological and medical studies involve recurrent events. For example, patients at risk for cerebrovascular events may experience transient ischemic attacks, patients with heart problem may have several myocardial infarctions, cancer patients may experience recurrent superficial tumors [3], repeated occurrence of asthma attacks in asthmatic patients, individuals with chronic disease may experience repeated episodes of hospitalization and so on. For each of the above examples, the event of interest differs, but nevertheless may occur more than once per subject. A logical objective for such kind of data is to assess the relationship of relevant predictors to the rate in which events are occurring, allowing for multiple events per subject [4]. Recurrent events are also observed in bladder cancer patients.

A bladder is a hollow organ in the lower abdomen (pelvis). It collects and stores urine produced by the kidneys. Cancer occurs when normal cells undergo a transformation whereby they grow and multiply without normal controls [5]. Bladder cancer, therefore, is a disease in which malignant cancer cells form in the tissues of the bladder. As more and more cells are produced, the tumor increases in size. According to National Cancer Institute, there are three types of bladder cancers that begin in cells in the lining of the bladder. These cancers are named for the type of cells that become malignant cancers: *Transitional Cell Carcinoma* (TCC): a cancer that begins in cells in the innermost tissue

layer of the bladder. These cells are able to stretch when the bladder is full and shrink when it is emptied. Most bladder cancers begin in the transitional cells. *Squamous cell carcinoma*: Cancer that begins in squamous cells, which are thin, flat cells that may form in the bladder after long-term infection or irritation and the third type is, *Adenocarcinoma*: cancer that begins in glandular (secretory) cells that may form in the bladder after long-term irritation and inflammation [6].

A cancer that is confined to the lining of the bladder is called superficial bladder cancer. In North America, South America, Europe, and Asia, the most common type of urothelial tumor diagnosed is transitional cell carcinoma; TCC constitutes more than 90% of bladder cancers in those regions [7].

The stage of a bladder cancer is a standard summary of how far the cancer has spread. It is one of the most important factors in selecting treatment options and estimating a person's outlook for recovery and survival (prognosis). The staging system normally used in bladder cancer is called TNM, which stands for 'Tumor' (size of the tumor), 'Node' (if it has spread to the nearby lymph nodes) and 'Metastasis' (if the cancer has spread to other parts of the body). There are usually four pathologic stages of bladder cancer: In *stage 0*, abnormal cells are found in tissue lining the inside of the bladder. These abnormal cells may become cancer and spread into nearby normal tissue. Stage 0 is divided into stage 0a (Papillary Carcinoma) and stage 0is (Carcinoma in Situ), depending on the type of the tumor. In *stage I*, cancer has formed and spread to the layer of tissue under the inner lining of the bladder. In *stage II*, cancer has spread to either the inner half or outer half of the muscle wall of the bladder. *Stage III* cancer has spread from the bladder to the fatty layer of tissue surrounding it and may have spread to the reproductive organs. *Stage IV*, cancer has spread from the bladder to the wall of the abdomen or pelvis [6, 8]. Figure 1 below shows the different stages in the bladder cancer [9].

In addition to stages, the grade of the bladder cancer provides important information and can help guide treatment. The tumor grade is based on the degree of abnormality observed in a microscopic evaluation of the tumor. The three bladder cancer grades are [10]:

- *Grade 1* cancers have cells that look very like normal cells - they are also known as 'low grade' or 'well differentiated' and tend to grow slowly and are not likely to spread
- *Grade 2* cancers have cells that look more abnormal - also called 'medium grade' or 'moderately differentiated' and may grow or spread more quickly than low grade.

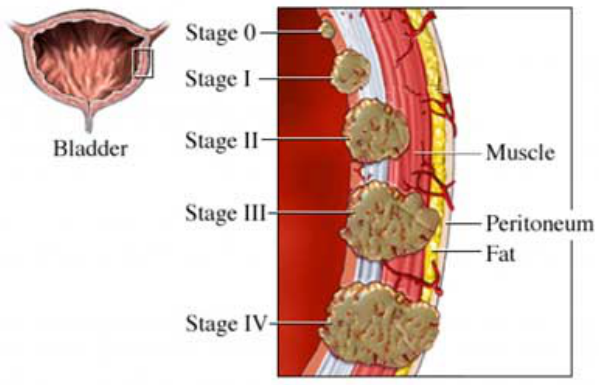

Figure 1: A diagram showing the different tumor stages in the bladder

- *Grade 3* cancers have cells that look very abnormal - are 'high grade' or 'poorly differentiated' and are more quickly growing and more likely to spread.

Smoking cigarettes, chemical exposures at work and diet are some of the factors that increase the person's risk of developing bladder cancer. Moreover, several studies support that males have the highest chance to develop the cancer than females. According to Steinberg and Katz, bladder cancer affects three times as many men as women. Women, however, often have more advanced tumors than men at the time of diagnosis. Whites, both men and women, develop bladder cancers twice as often as other ethnic groups. Bladder cancer can occur at any age, but it is most common in people older than 50 years of age. The average age at the time of diagnosis is in the 60s. However, it clearly appears to be a disease of aging, with people in their 80s and 90s developing bladder cancer as well [5].
Bladder cancer has the highest recurrence rate of any malignancy. Although most patients with bladder cancer can be treated with organ-sparing therapy, most experience either recurrence or progression, creating a great need for accurate and diligent surveillance. Due to its high recurrence and progression rate, bladder cancer is the most expensive cancer to treat on a per patient basis [5, 7].

1.2 Objective

The main objective of this study is to compare different marginal and random effect (conditional frailty) models to provide absolute risk prediction of recurrent events in bladder cancer patients. The report also focuses on validating the predictive ability of the models for future predictions.

1.3 Organization of the thesis

This report is partitioned into several sections. Under the methods and materials section, a brief description of the dataset in *Section 2.1* and statistical methods used to fit the recurrent events in *Section 2.2-2.3* are provided. Different marginal and random effect model (frailty model) are discussed under *Section 2.3.1* and *2.3.2* respectively. *Section 2.4* deals with how to handle missing covariates in the dataset and *Section 2.5* methods of evaluating the prediction models are discussed. Results from marginal models and random effect models are presented under *Section 3.2* and *Section 3.3* respectively. Finally, discussion and conclusions are provided in *Section 4*.

2 Methods and Materials

2.1 Data Description

The dataset employed under this study is recurrent event data collected from a bladder cancer study conducted at different hospitals in Rotterdam with followed-up period between the years 1974 and 2011. In this study, all patients had primary bladder tumors when entering the study and these tumors were removed. A total of 615 patents, were followed during the study and 1280 observations were recorded. Most patients had multiple recurrences of tumors during the study. The new tumors were removed at each visit. The outcomes of interest were recurrences of new cancer tumors during the follow-up periods and the time to the recurrence event. For patients with multiple recurrences of the tumor, times between successive recurrences were recorded in days (Time). Furthermore, for each patient, some potentially important demographical and clinical covariates were recorded including: gender (Sex), age at time of operation (Age), tumor size groups ($\leq$3cm, 3-5cm and $\geq$ 5cm), tumor multiplicity (solitary or multiple tumors), pathologic stage (Stage), pathologic grade (Grade), Metastasis (yes or no) and presence of Carcinoma in Situ (CIS). Every patient has a unique identification number (Pat-ID), where each event within a patient was ordered by an indicator (Sequence). Another important covariate, number of previous recurrences at each event, was also included in the analysis. An indicator variable, Event, represents the status of every observation, "1" if tumor recurrence observed and "0" if the event is censored due to several reasons. For simplicity it is assumed that the censoring mechanism was independent of the recurrent event process in this study.

2.2 Modeling Recurrent Events with Cox-PH Models

2.2.1 Ordinary Cox Model

The Cox proportional hazard model has become by a wide margin the most used procedure for modeling the relationship of covariates to the survival or other censored outcomes. The mathematical notations of the standard Cox model can be briefly illustrated as: Let $X_{ij}(t)$ be the j^{th} covariate of the i^{th} subject, where i = 1, . . . , n and j = 1, . . . , p and X_i denotes the covariate vector for subject i. The Cox model specifies the hazard for individual i as:

$$\lambda_i(t) = \lambda_0 exp(X_i(t)\beta) \tag{1}$$

where λ_0 is an unspecified nonnegative function (the baseline hazard), and β is a vector of coefficients. When the hazard ratio of pair of subjects with fixed covariate vectors X_i is constant over time, the model is known as proportional hazard model. For untied failure time data, Cox in 1972, proposed the estimation of β based on the partial likelihood function [11]:

$$PL(\beta) = \prod_{i=1}^{n} \prod_{t:Y_i(t)=1} \left\{ \frac{exp(X_i(t)\beta)}{\sum_j Y_j(t) exp(X_j(t)\beta)} \right\}^{dN_i(t)} \tag{2}$$

where $Y_j(t)$ is equal to 1 if the j^{th} subject is at risk of a failure event just prior to time t, and is equal to 0 otherwise. $N_i(t)$ is the number of observed failures for subject i and $dN_i(t)$ denotes the increment in $N_i(t)$ over the time interval $[t, t + dt)$. Differentiating the log partial likelihood with respect to β gives the score equation, $U(\beta)$, and the maximum partial likelihood estimator for β is obtained by solving the score equation, $U(\beta) = 0$. The solution $\hat{\beta}$, is consistent and asymptotically normally distributed with mean β, the true parameter vector and variance $I^{-1}(\hat{\beta})$, the inverse information matrix.

2.3 Cox Models for Multivariate Failure Events

Multivariate failure time data arise when each study subject can potentially experience several events. There is an increasing interest and need to apply survival analysis to datasets with multiple events per subject. This includes both the cases of multiple events of the same type and events of different types [9]. The analysis of multiple events per subject cannot be approached by a standard Cox model, where the assumption of independence of observations is not valid. In order to account for intra-subject correlation, extensions of the above Cox proportional models are used [12]. Apart from the major reason to extending the Cox model (i.e. the intra-subject correlation), there are other

concerns such as: discontinuous intervals of risk, stratum (several recurrences) by covariate interactions and the structure of the risk sets [11, 12]. The primary difference in the Cox model used for analyzing recurrent event data versus nonrecurrent (one time interval per subject) data is the way several time intervals on the same subject are treated in the formulation of the likelihood function maximized for the Cox model [4]. i.e. for recurrent survival data, a subject with more than one time interval remains in the risk set until his/her last interval, after which the subject is removed from the risk set. In contrast, for single event data, each subject is removed from the risk set at the time of failure or censorship.

Several approaches to handling multiple recurrence events data have been proposed in literatures [11, 12, 13, 14]. The multiple events data can be further classified into ordered and unordered data. For ordered data, there is a natural ordering of the multiple failures within a subject, which includes recurrent events data as a special case [14]. To analyze such types of data, many researchers estimate and compare event rates using the chi-square test or event times until the occurrence of the first event, or the overall event time using the Cox model. Such conventional methods are inefficient because they use only parts of the available information in the data [15].

2.3.1 Marginal Models

The three most common marginal model approaches for analyzing ordered outcomes (i.e. multiple events of the same type) are: the independent increment, *Andersen-Gill*, (AG) [16], the *marginal*, Wei, Lin, and Weissfeld (WLW) [17] and the *conditional*, Prentice, Williams and Petersen (PWP) models [18]. These methods offer great flexibility in the formulation of strata and risk sets, and have a well-developed estimator of variance. All the three are "marginal" regression models in that β is determined from a fit that ignores the correlation between the events followed by a correction of the variance. The main difference between these models is the definition of risk sets and stratification of the baseline hazard function. The three models need different time intervals and strata representations in defining the risk sets. A hypothetical subject having two events and censored at the last time interval under each model is shown in Table 1.

Table 1: Risk set representation of a hypothetical subject for marginal models

	Interval	Event	Stratum
Counting Process,(AG)	(0,10]	1	1
	(10, 30]	1	1
	(30, 42]	0	1
Conditional (PWP)	(0,10]	1	1
	(0, 20]	1	2
	(0, 12]	0	3
Marginal (WLW)	(0,10]	1	1
	(0, 30]	1	2
	(0, 42]	0	3

The models allow for population average estimation of covariate effect and the analysis of these models is based on the following three steps [11, 12]:

- Decide on a model (issues such as covariate selection, inclusion of strata, etc.) and structure the data set accordingly.
- Fit the data as an ordinary Cox model, ignoring the possible intra-subject correlation (i.e. treating multiple events from a subject as independent).
- Replace the standard variance estimate with one which is corrected for the possible correlations.

The marginal models can overcome this assumption of independence for the estimation of the variance of β by an appropriate correction based on a grouped jackknife estimate [4, 11, 15]. Grouped jackknife values are defined as $J_i = \beta - \beta_{(i)}$, where $\beta_{(i)}$ is the result of the fit that includes all of the individuals except individual i. It is denominated as grouped because in the multiple event case, one individual contributes several observations, and removing a subject implies removing a group of observations. Full description of how to compute the grouped jackknife values directly from Newton-Raphson iteration method is found in [11]. The computed variance can be viewed as sandwich estimators and the resulting robust covariance matrix estimator is given by $V_R = I^{-1}(U)I^{-1}$. These sandwich estimates are familiar from robust variance estimation in generalized estimating equations (GEE) proposed by Liang and Zeger [19]. This grouped jackknife estimate is typically more variable than the ordinary variance of the Cox model, since it is a robust variance estimate that adequately addresses the repeated event correlation. In the next sections, brief descriptions and differences of the three marginal approaches are discussed.

I. The Andersen-Gill Models

This model, also known as independent increment model, is a derivation of the Cox proportional hazard model as a counting process [14, 18]. The method is the simplest, but makes the strongest assumptions. Each subject is represented as a series of observations (rows of data) with time intervals as: (entry time, first event], (first event, second event],..., (m^{th} event, last follow up]. The intensity process for subject i is:

$$Y_i(t)\lambda_0(t)exp(X_i(t)\beta) \tag{3}$$

The difference with the standard Cox model lies in the definition of the at-risk indicator $Y_i(t)$. For survival data, the individual ceases to be at risk when an event occurs and $Y_i(t)$ takes value zero, but for the Andersen-Gill model for recurrent events, $Y_i(t)$ remains one as events occur, i.e. individuals are assumed to be under risk for all their events until censoring. Under this model, the risk of a recurrent event for a patient follows the usual proportional hazards assumption and is unaffected by earlier events that occurred to the patient unless terms that capture such dependence are included explicitly in the model as covariate [16, 20]. No extra strata or strata by covariate interaction terms are introduced for the multiple events. The Andersen-Gill formulation of the Cox proportional hazards model has a number of advantages, including the ability to accommodate left-censored data, time-varying covariates, multiple events and discontinuous intervals of risks.

II. The marginal or WLW Models

Wei, Lin and Weissfeld proposed a marginal approach to the analysis of multivariate failure time data. In this model, the ordered outcome dataset is treated as if it were an unordered competing risk case. The number of strata in the analysis will be equal to the maximum number of events a patient reports in the study. Every subject will have one observation in each stratum. The hazard function for the j^{th} event for subject i is:

$$Y_{ij}(t)\lambda_{0j}(t)exp(X_i(t)\beta_j) \tag{4}$$

In most applications the analysis has been on the "time from study entry" scale, since all the intervals start from zero the model can in this case be fit without recourse to the counting process style of input [15]. Unlike the AG model, this model allows a separate underlying hazard for each event and for strata by covariate interactions, as shown by the notation β_j. In the WLW model the at-risk indicator for the j^{th} event, $Y_{ij}(t)$, is one until the occurrence of the j^{th} event, unless, of course, some external event causes censoring. When either of those occurs, it becomes zero, indicating that subject is no longer at risk after the last given event [11, 21].

III. The Conditional or PWP Models

The model is proposed by Prentice, Williams and Petersen. This model clearly defines the order of the events. The model is known as conditional because, a subject is not at risk for the k^{th} event until he/she has experienced event $k-1$. Like the Andersen-Gill model, the counting process time scale (*conditional I model*) or time between successive events, i.e. the "gap" time scales (*conditional II model*) can be used. Both the conditional models are similar in formulation of risk set but they only differ in the time scale they used in fitting the model. However, unlike Andersen-Gill model, each event is assigned to a separate stratum. The use of time-depending strata means that the underlying hazard function may vary from event to event, unlike the AG model, which assumes that all events are identical [11, 12, 17]. The hazard function for the j^{th} event for subject i is:

$$Y_{ij}(t)\lambda_{0j}(t)exp(X_i(t)\beta_j) \tag{5}$$

The primary difference between the WLW and PWP models is in the definition of the at-risk indicator and the definition of the strata in the analysis. In the PWP model the at-risk indicator, $Y_{ij}(t)$, is defined as 0 until the $j-1^{st}$ event and only then becomes one. Once the j^{th} event occurs, $Y_{ij}(t)$ becomes 0 again. The PWP model can be seen as a stratified AG model with event-specific baseline hazards and a restricted risk set.

In summary, the schematic form of the risk set formation in the different marginal models is presented in Figure 2. Each arrow represents a stratum or an event.

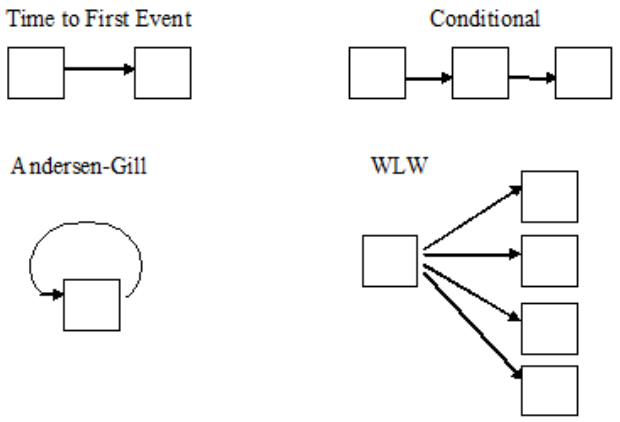

Figure 2: Schematic form of the three marginal models

In general, the AG model and the PWP model can be used in the analysis of repeated failure outcomes of the same type, while the approach by the WLW model can be applied to both multiple events of the same type and different types as long as there is not a predetermined ordering. The WLW model has a semi-restricted risk set that allows subjects to be at risk for as many events as the maximum number of events reported per subject in the study, even if most of the subjects only had one event. However, this leads to overestimation of the covariate effects [12, 22]. When the model is correctly specified (no important covariates are omitted) the PWP model and the AG model estimate unbiased covariate effect and require similar sample size to obtain the same precision in the estimation, while the WLW model estimates biased covariate effect and requires a larger sample size. The PWP model and the AG model are considered to be more efficient than the WLW model [11]. The most appropriate modeling strategy should be chosen based on the type and nature of the multiple events structure [21]. Apart from marginal models, random effect models are widely used in modeling multiple failure time data, including recurrent events. The frailty model is a well known random effect model and in the next section a brief description of this method is provided.

2.3.2 Random Effect Models: The Frailty Model

An important issue in analysis of recurrent events data is how to account for the dependence of the events in an individual and any unobserved heterogeneity of the event propensity across individuals [23]. This is due to the fact that the recurrent events of an individual are possibly correlated because of underlying characteristics of the individual. The heterogeneity of the event rates across individuals may not be fully reflected in the measured baseline variables [24]. The marginal models, discussed above, are variance-correction models since they all do not require specification of the magnitude of the correlation of the recurrent events in an individual. They simply adjust for this correlation by using a robust variance estimator. In contrast to the marginal models, frailty or random effects models incorporate heterogeneity and event dependency into the estimator by making assumptions about the frailty distribution and incorporating it into the model estimates and thus present a more promising alternative [24]. The underlying logic of frailty models is that some subjects are intrinsically more or less prone to experiencing the event of interest than are others, and that the distribution of these effects can be at least approximated. i.e. individuals have different frailties, and that those who are most frail will die earlier than the others [24, 25].

Several complex frailty models are available. In this paper, the simplest model, the shared frailty model is considered due to its simplicity. In this model, all the records within each

patient share a common frailty, each unit belongs to precisely one category, and frailties of different categories are independent [13, 24, 25]. In this particular bladder cancer data, the common frailty is shared by the recurrent events of an individual. Thus, the shared frailty model is given by:

$$h_i(t) = h_0(t)exp(\beta^T X_i(t) + w_i) \tag{6}$$

where $h_i(.)$ is the hazard function for the i^{th} individual; $h_0(.)$ is the baseline hazard, β is the fixed effects vector, X_i is the vector of covariates and w_i is a vector containing the unknown random effects or frailties for the i^{th} subject [25]. The w_i's, i=1,..., n are independent and identically distributed sample from a density $f_W(.)$. The frailty model can be rewritten as:

$$h_i(t) = h_0(t)u_i exp(\beta^T X_i(t)) \tag{7}$$

where $u_i = exp(w_i)$ is known as the frailty for the i^{th} subject. Model (6) and (7) are conditional hazard models given the u_i's. Furthermore, distributional assumptions has to be made for the frailty and usually are chosen from the class of positive distributions; in applied work, the most widely used are the gamma and Gaussian distributions. However, the one-parameter gamma distribution with mean 1 and variance θ frailty is being by far the most frequent due to the flexibility of the distribution and is represented as [24-26]:

$$f_U(u) = \frac{u^{\frac{1}{\theta}-1}exp(-\frac{u}{\theta})}{\theta^{\frac{1}{\theta}}\Gamma(\frac{1}{\theta})} \tag{8}$$

with $\Gamma(.)$ represents the gamma function. This gives the following interpretation: individuals with $u_i > 1$ ($u_i \leq 1$) are frail (strong) (higher risk, lower risk, respectively). The parameter θ provides information on the variability (the heterogeneity) in the population of groups [27, 28]. In a semi-parametric approach, the Cox regression models assume that the baseline hazard is unspecified. Furthermore, the frailties (w_i) are viewed as unobservable data. Under the idea of shared gamma frailty model, it is difficult to maximize the likelihood to estimate the parameters (β, θ) as it contains the unobserved frailty term and the unspecified hazard [25, 28]. However, such kind of problem of estimation can be approached by Expectation-Maximization (EM) algorithm. Though, this algorithm is slow, computationally intense and the implementation has not appeared in any of the widely used available packages [11]. Instead, another approach resulting with similar estimation of the parameters is through adaptation of the partial likelihood approach for the Cox model known as the penalized Cox partial likelihood method. Since the frailties are assumed to be a random sample from the frailty density, a penalty term is subtracted from the partial likelihood in order to account for that. This method of estimation is easier to implement. The details how the EM algorithm and penalized partial likelihood approaches work are discussed in [25, 26].

2.4 Handling Missing Covariates

The construction of a prognostic or prediction models ideally requires a large database with complete information on all potential prognostic factors [29]. However, often missing covariate data occur and complicates analysis. These missing covariates may introduce bias and lead to misleading conclusions if handled inappropriately. Many strategies for handling missing covariates when fitting a Cox proportional hazards model have been proposed [30, 31, 32]. These include:

- simple deletion approaches of complete case analysis, where only the cases with complete data for all collected variables are analyzed,
- available case analysis, where the cases with complete data for the variables in the fitted model are analyzed utilizing the largest possible data set,
- variable omission, where the incomplete variable is excluded from the model.
- utilizing all data, include analyzing the missing data as a separate category (i.e create extra level for missingness for categorical variables)
- single imputation, in which a single value is substituted for each missing value and
- multiple imputation (MI), where more than one independently completed data sets are obtained.

In this bladder cancer dataset, three categorical variables (tumor size, multiplicity and CIS) have missing values where tumor size has the highest percentage of missing (62%). The most commonly used approaches for dealing with missing data are complete case analysis and available cases. However, these methods were not suitable in this context because only 37% of the observations were complete, which might result in great loss of power and biased estimates. Omission of the important incomplete covariates may also result in poor model prediction ability. The other approach is including the missing data as a separate category, even though this method uses all the cases in the data, it is not recommended in several literatures since the method almost always results in biased estimates and inefficiency [33, 34]. Single and multiple imputation, can provide unbiased estimates under the assumption of missing at random (MAR). This means that the probability of missingness does not depend on unobserved information. However, single imputation method commonly results in underestimation of standard errors since it imputes all the missings only once and does not incorporate uncertainty.

Thus, multiple imputation is the most widely used technique to draw valid statistical inference in the presence of missing data. However, the assumption of missing at random is

not commonly testable. In MI, where each missing value is replaced with a set of m (>1) independent values to give m separate complete datasets, incorporates uncertainty of the missing data that cannot be achieved with single imputation ($m = 1$). The m completed datasets are analyzed individually using standard statistical methods and the results combined into one summary estimate. The parameter estimates of interest are averaged and a variance estimate is obtained that incorporates both the within and between imputation variability [33]. In this report, MICE (multiple imputation by chained equations) technique in R package with m=5 was used in each model to handle the missingness observed in the three covariates.

2.5 Evaluation of Prediction Models

When the purpose of a survival regression model is to predict future outcomes, the predictive accuracy of the model needs to be evaluated before practical application [35]. There are various ways to assess the performance of statistical prediction models. The usual statistical approach is to quantify how close predictions are to the actual outcomes using explained variation (R^2) and goodness-of-fit statistics [36, 37]. *Discrimination* is the most commonly applicable approach that quantifies the ability of the model to correctly classify subjects into subgroups in which each individual belongs [38, 39]. A C- (concordance) statistic has been widely used to assess and compare prediction models with respect to their ability to discriminate individual risks.

The C-index is defined as the proportion of all usable patient pairs in which the predictions and the outcome are concordant. It measures predictive information derived from the set of predictor variables in the model [38]. Usable patient pairs means that, pairs should be either event *vs* event or event *vs* censored, but not censored *vs* censored, these pairs are unusable pairs in calculating the C-index. For Cox proportional model, given comparable or usable pairs (i, j), if $T_i > T_j$ & $\beta^T x_i > \beta^T x_j$ then (i, j) is concordant pair where as if $T_i > T_j$ & $\beta^T x_i < \beta^T x_j$ then (i, j) is disconcordant pair [40, 41]. Thus, the Harrell$'s$ C index is computed as ratio of number of concordant pairs and number of comparable pairs [42]. C-statistics of the model can also be computed from the Somers$'$ D_{xy} rank correlation for censored response variable since C and D_{xy} are related by: $D_{xy} = 2 * (c - 0.5)$ [40]. A concordance probability of 1.0 represents a model that has perfect discrimination, whereas a value of 0.5 indicates that the predictions are no better than chance [43, 44]. In this report, the discriminative ability of the marginal and random-effect models for the first four recurrent events were assessed by computing their concordance probabilities.

2.6 Model Development and Diagnosis

The key of fitting any of the three marginal models for multiple events data is the creation of an appropriate data set. The dataset were arranged accordingly for each model. The "counting process" data input for Andersen-Gill and conditional I models, the "gap" time scale for the conditional II model and the "time from study entry" scale for marginal (WLW) model were used in the analysis. The assumption of non-informative censoring, that is, patient′s follow-up time is independent of his/her event times is considered throughout the report. Further, in the presence of tied events, Efron correction method for ties were applied. For simplicity, patients that have no events and 0 days of follow-up are removed from the analysis since they add nothing to the likelihood. The covariate stage contains only one observation in its fourth level (stage IV), in the statistical analysis, it is merged to the stage II. The continuous variable age was centered at 65, ease of computation of the baseline hazard rate. The covariates gender (female), age, tumor size (<3 cm), tumor multiplicity (solitary), stage (0), grade (G1) and carcinoma in situ (no) with possible interactions were included in the full model, where levels in the bracket are taken as reference group for corresponding covariates. Variable reductions were performed using manual backward selection at 5% level of significance. However, clinically important variables are kept in the model even though they are not statistically significant. In all analysis, multiple imputation with five independently completed datasets were analyzed and results are averaged out.

Apart from all the variables listed above, the number of previous recurrence (# of prev.rec.) was calculated as one minus the number of recurrence (*sequence*) at different time intervals for each patient. This covariate was included only in the Andersen-Gill model to see the effect of previous recurrences on the future recurrences. This covariate was not included in the other marginal models since the ordering of the recurrence events were accounted for the variable sequence or stratum number at each time interval.

After fitting the models, it is reasonable to test the underlying assumption of the proportionality of the Cox model for valid interpretation. Possible violation of proportional hazard assumptions were checked formally and graphically by looking the trends of Schoenfeld residuals against time for each predictor separately. This assumption was checked for all the models considered in this report.

2.7 Software and Tools

All the statistical analysis are carried out using SAS version 9.2 and R version 2.13.1 statistical packages. The *survival, Design* and *mice* libraries from R package are used to fit the Cox models, to compute the concordance probability and to perform multiple imputation respectively. Alpha= 5% level of significant is used throughout the report.

3 Results

3.1 Exploratory Data Analysis

A total of 615 patients were included in the analysis and the main characteristics of patient at entry level and at different recurrent events were summarized in Table 2. Patient and tumor characteristics at time to first recurrence are summarized as follows: the average age of the patients is 65 years with standard deviation of 12.2. From total patients, 495 (almost 80%) are male patients and only 20% are females. In 193(68%) of the patients, the maximum diameter of the tumor is less than 3 cm, 54(19%) had 3-5 cm and 38(13%) had greater than 5 cm tumor diameters. However, for 330(54%) of the patients, the tumor size is missing. More patients have solitary tumor 313(68%) than multiple tumors (37%). For 119 patients, the tumor multiplicity are not measured. Moreover, 62% of the patients are in stage 0 and only one patient in stage IV. More patients (41%) are in grade 2 and only 4% had carcinoma in situ. Similarly, the distribution of the covariates on the second, third and fourth recurrence are presented in Table 2.

Individual patients experienced from zero to fifteen recurrences during follow-up. Out of 615 patients, 338 (55%) did not show any recurrence during the follow-up period. The bladder cancer recurred at least once in 277 patients (Table 3) and the highest recurrence events (14 and 15 times) were observed in two patients. Moreover, Table 3 shows a summary of follow-up times and number of patients with or without event for the first five consecutive recurrent events. The median follow-up time to the first recurrent event was 24 months and starts decreasing for the higher subsequent recurrent events.

Table 2: Patient characteristics in #(%) at different recurrent events

	$1^{st}event$	$2^{nd}event$	$3^{rd}event$	$4^{th}event$
Gender:				
Male	495(80.5)	207(75.5)	110(74.8)	58(73.4)
Female	120(19.5)	67(24.5)	37(25.2)	21(26.6)
Age:				
<65	338(55.0)	143(52.2)	86(58.5)	45(57.0)
≥65	277(45.0)	131(47.8)	61(41.5)	34(43.0)
Tumor-size group:				
<3cm	193(67.7)	70(89.7)	39(86.7)	25(92.6)
3-5cm	54(19.0)	5(6.4)	3(6.7)	1(3.7)
≥5cm	38(13.3)	3(1.1)	3(6.7)	1(3.7)
Missing	330	196	102	52
Tumor multiplicity:				
Solitary	313(63.1)	97(43.3)	44(36.1)	24(33.8)
Multiple	183(36.9)	127(56.7)	78(63.9)	47(66.2)
Missing	119	50	25	8
Pathologic stage:				
0	384(62.4)	182(66.4)	112(75.7)	64(81.0)
I	208(33.8)	58(21.2)	16(11.5)	12(15.2)
II	22(3.6)	34(12.4)	19(12.8)	3(3.8)
IV	1(0.1)			
Pathologic Grade:				
G1	171(27.8)	84(31.0)	49(33.3)	29(36.7)
G2	252(41.0)	113(41.2)	58(39.5)	39(49.4)
G3	192(31.2)	77(28.1)	40(27.2)	11(13.9)
Carcinoma in situ:				
No	540(96.0)	240(91.30)	132(95.0)	73(93.6)
Yes	23(4.0)	23(8.8)	7(5.0)	5(6.4)
Missing	52	11	8	1

Table 3: Summary of time between consecutive recurrent events

	Follow-up time(months)			# of patients with		
	Min	Max	Median	event	censored	Total
1^{st} recurrence	0.13	303.3	24.2	277	338	615
2^{nd} recurrence	0.23	276.2	17.3	147	127	274
3^{rd} recurrence	0.63	244.2	15.4	79	68	147
4^{th} recurrence	5.3	313.1	13.3	45	34	79
5^{th} recurrence	1.4	220	13.4	33	12	45

As observed in Table 2 above, for three covariates some observations were not measured or available. The missimgness of the data with respect to the covariates were explored for further data analysis purpose. Below Table 4 shows the missing patterns for the covariates used in the models. Out of 1280 records, only 471(37%) of the data were complete and lots of missing values were observed on the covariate tumor size (62%).

Table 4: Missing data pattern

Group	Sex	Age	Stage	Grade	Tumor size	Multiplicity	CIS	Freq.(%)
1	O	O	O	O	O	O	O	471 (36.8)
2	O	O	O	O	O	O	M	15(1.2)
3	O	O	O	O	O	M	O	3(0.2)
4	O	O	O	O	O	M	M	1(0.1)
5	O	O	O	O	M	O	O	559(43.7)
6	O	O	O	O	M	O	M	23(1.8)
7	O	O	O	O	M	M	O	171(13.4)
8	O	O	O	O	M	M	M	37(2.9)
% missing out of total incomplete cases					62	17	6	

O= Observed, M=Missing

Before modeling the multiple failure times for the recurrent events, it is important to see the variation in the recurrence pattern between male and female patients in each of the first four failure events without adjusting for the other covariates. Figure 3 presents the cumulative hazards for the first four consecutive events by gender. It clearly suggests that the risk of a new event does not remain constant. It is observed that females have higher risk than males in the first recurrence, while no clear difference is obsered in the other events. However, such plots totally ignore the dependency of the recurrent events. The dependency of the events and other important covariates are investigated in the next section of marginal and random-effect model formulations.

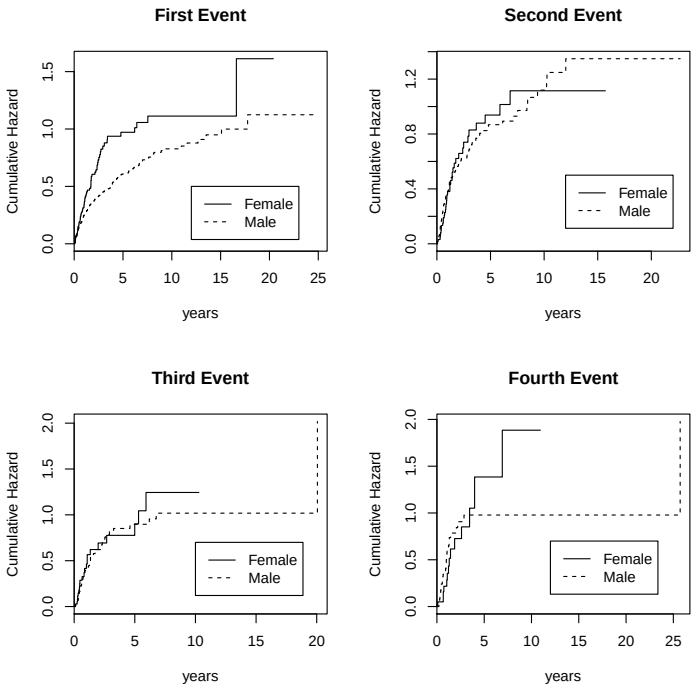

Figure 3: Cumulative hazard for the first four consecutive events by gender

3.2 Marginal Models

The Andersen-Gill model which assumes that the data are a set of independence increment was fitted to model the multiple recurrent events. The independent assumption here means that each recurrence event is not conditioned by previous recurrences. Univariate analysis based on Andersen-Gill model was carried out (result not presented) and except the covariates CIS and age, all were identified as important prognostic factors. Further, all the covariates were included in the multivariate analysis to see the effect of the factors collectively. The functional form of the continuous variable age was assessed including restricted cubic spline function but models were not improved and only the linear relationship was kept in the model. Results of the final model fit after possible reduction of some non-significant covariates were presented in Table 5. Since the model does not allow strata or strata by covariate interactions, the hazard functions have constant effects across all events. The model-based standard errors (not presented here) are smaller than the robust standard errors; this is due to the fact that the robust standard errors account for the dependency between recurrence events coming from the same patient. The results from this analysis indicate that the number of previous recurrence, gender, tumor mul-

tiplicity and stage have statistically significant effects on the recurrence of the bladder cancer.

Table 5: Parameter estimates and robust standard errors for marginal models

	AG	Conditional I	Conditional II	Marginal
Male	-0.242(0.097)*	-0.166(0.090)*	-0.128(0.091)	-0.463(0.187)*
Age	-0.003(0.004)	-0.001 (0.003)	-0.004(0.003)	-0.001(0.005)
Multiplicity	0.475(0.105)*	0.424(0.103)*	0.458(0.096)*	0.788(0.127)*
Tumor size (3-5cm)	-0.131(0.170)	0.004(0.169)	0.005(0.170)	-0.356(0.237)
Tumor size ($\geq$5cm)	0.092(0.238)	0.187(0.243)	0.198(0.231)	-0.177(0.205)
Stage I	-0.231(0.138)	-0.220(0.136)	-0.125(0.120)	-0.467(0.172)*
Stage II	-1.249(0.268)*	-1.328(0.270)*	-1.198(0.263)*	-1.356(0.314)*
Grade 2	0.069(0.111)	0.030(0.104)	0.071(0.093)	0.126(0.137)
Grade 3	0.210(0.161)	0.180(0.152)	0.219(0.132)	0.252(0.212)
# of prev.rec	0.269(0.017)*			

* is significant at 5% level of significance

Since the Andersen-Gill model is based on a strong assumption of independence, a second conditional model, which assumes that a patient cannot be at risk for a subsequent recurrent events without having experiencing the previous event, was fitted. First conditional model (conditional I) which takes the same time format as the Andersen-Gill model and second conditional model (conditional II) with gap time were fitted. conditional II model is very useful for modeling the time between each of the recurring event rather than the full time course of the recurrent event process. The final models based on conditional I and conditional II models are displayed in Table 5. The models have common covariate effect across the events, since time-depending strata was not incorporated. This means that the underlying hazard function is assumed to be the same from event to event. Unlike the Andersen-Gill model, both the conditional models account for the order of the recurrent events properly in the model formulation. As observed from Table 5, the same sets of covariates were found statistically significant with slight differences in their parameter estimates.

The fourth model considered is a marginal (WPW) which assumes every subject to be at risk as the maximum number of recurrent events occurred in the study (k=16) even if a patient has one recurrence. i.e., every subject has 16 observations, one in each stratum. This model uses different time scale in that, the time for each event starts at the beginning of follow up (zero) for each patient. It considers each event separately and models all the available data for the specific event. The last column of Table 5 shows the parameter estimates and corresponding robust standard errors for this model. The

parameter estimates (in absolute value) and the robust standard errors obtained from this model are larger compared to the estimates obtained from the AG and PWP models. That is, the WPW model overestimates the covariate effects due to the fact that every subject has as many records as the maximum number of event occurred in the data.

It is necessary to investigate the hazard rates in the different recurrent events since covariates might have different effects across the events. For this reason, further the conditional (PWP) and marginal models were fitted again but considered here with time-depending strata. That is, covariates or hazard functions have event specific interpretations. Due to the small numbers of participants who had more than four recurrent events in the study, event-specific analysis for both PWP and WLW models were done by limiting the maximum number of recurrences to four (k= 4). In such kind of analysis, the estimates in the first stratum are exactly the same as time to the first recurrent analysis. Estimates from the second stratum corresponds to the analysis of time to the second recurrence and so on (result not presented here). Here, it is important to test a linear hypothesis that checks whether covariates have identical effects across the four strata (interactions between the strata and each covariates). This null hypothesis can be stated as:

$$H_0 : \beta_{i1} = \beta_{i2} = \beta_{i3} = \beta_{i4} = \beta$$

Where, β_{i1} is the effect of covariate i on the first stratum. If these tests for identical effect are significant then the hazard function has different effect on each event. However, the test indicates that no evidence of heterogeneity in the effect risk factors across the recurrent event (p-value were greater than 0.12 in all cases). However, the cumulative hazard functions that correspond to the fit of the stratified conditional model can reveal important features of the data. The resulting plot, presented in Figure 4, shows that the time to the first recurrent event has slightly different (smaller) cumulative hazard than the other gap times. Second, third and fourth recurrent events do not show clear differences in their cumulative risks, supporting the hypothesis tested above. Higher recurrence events and hazard function after 15 years were not presented in the plot since very few information are present in each event and after this period.

3.2.1 Comparison of Marginal Models

Even though the Andersen-Gill model is easy to formulate and interpret, it is based on the strong assumption that multiple event times for an individual are mutually independent. To capture such dependence, the variable, number of previous recurrence, was included in the models as covariate. Inclusion of the event strata (order of the events) in conditional I and II models did not show great changes in the parameter estimates as compared to

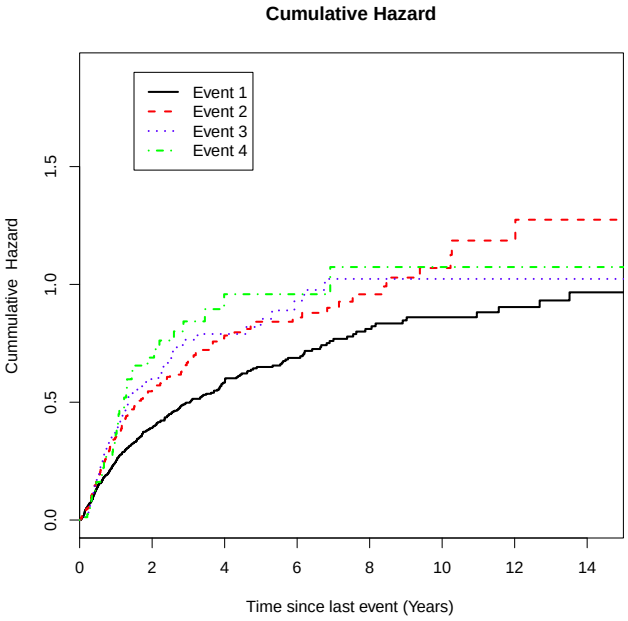

Figure 4: Cumulative hazard for the first four events

Andersen-Gill models. The Andersen-Gill and conditional I models enable to model the full time course of the recurrent event processes while conditional II models only the time between successive event. Furthermore, the Andersen-Gill and PWP model estimates are unbiased and more efficient than WPW model. For instance, the estimated hazard ratio ($exp(\beta)$) for *gender* is: *0.79*, *0.85*, *0.88* and *0.63* from AG, conditional I, conditional II and marginal models, respectively. The hazard ratio for the covariate differs somewhat for each of the four approaches, with the marginal model giving a much different result from those obtained in the other three approaches.

The global test for proportionality assumption for the four models were assessed. Results show that the tests were not significant in all the models except the marginal model, indicating that the PH assumption is satisfied. A sample plots of Schoenfeld residuals against each predictor variables for Andersen-Gill model (Figure A1 in appendix) show constant trend along the horizontal line around zero. The two conditional models are appropriate to account for the ordering of events properly and to see the effect of the covariates on the different strata (event).

In addition, the concordance probabilities for the first four recurrence events were computed separately and presented in Table 6. These concordance probabilities are used here to evaluate the discriminatory power and the predictive accuracy of the marginal models. Since, the C-statistics are estimated as a function of the regression parameters of the cor-

responding models and the covariate distribution only, they are asymptotically unbiased. However, when the parameter estimates are biased, the estimated C-statistics will be biased. As observed in Table 6, the discriminatory power of the Andersen-Gill, conditional I

Table 6: Concordance probability per recurrence event

	AG	Conditional I	Conditional II	Marginal	Frailty-gap	Frailty-CP
1^{st} recurrence	0.600	0.596	0.605	0.585	0.822	0.840
2^{nd} recurrence	0.562	0.565	0.565	0.638	0.732	0.735
3^{rd} recurrence	0.582	0.587	0.583	0.699	0.741	0.738
4^{th} recurrence	0.543	0.541	0.554	0.748	0.748	0.741

and II models are almost similar for each event. For instance, based on the Andersen-Gill model, the concordance probability can be interpreted as; for randomly selected pair of subjects having the first recurrence, the predicted survival and the observed survival are concordant with probability of 0.6. For the marginal model, the C-statistics is higher for the 2^{nd}, 3^{rd} and 4^{th} recurrences, this is because the estimated parameters are higher for this model as observed from Table 5.

In summary, the Andersen-Gill and the two conditional models have comparable predictive accuracy and discriminatory power. Moreover, the proportionality hazard assumption was satisfied in these models. While the marginal model has biased (overestimated) results. The Andersen-Gill model has important additional covariate (number of previous recurrence) and easy to interpret, thus, further interpretation and prediction can be done using this model. The model is also the best choice since the recurrent events are identical and event order does not bring great change.

3.3 Random-Effect Models: Frailty Model

All the marginal models fitted so far are handling the correlation between recurrent events from the same patient by only correcting the variance. In order to account for the dependency between the recurrent events or unobserved heterogeneity among patients properly, a semi-parametric random-effect, frailty model is fitted. Under these models, it is assumed that, conditional on some unobserved frailty random variable, the time to recurrence of tumor would follow a Cox proportional hazards model. Multiple recurrence events from each individual share a common frailty random variable, which accounts for the within-individual correlation. A counting process and gap time scale were used to fit the frailty models and the models are compared. Depending on the timescale selected, the interpretation of the time evolution will be entirely different in the different frailty models.

The parameter estimates, standard errors for the frailty models with counting process and gap time were presented in Table 7. As observed from the table, different parameter estimates for the covariates and variance of the frailty term were obtained from the two models. The variance of the frailty term (θ) was estimated to be 0.445 and 0.719 under the gap and counting process time scales, respectively. *Gender*, *multiplicity* and *stage* were found statistically significant in both models but the *number of previous recurrent* was significant only in the counting process frailty model.

Table 7: Parameter estimates for frailty models

	Gap-time	CP^{+}- time
Male	-0.256(0.121)*	-0.312(0.134)*
Age	-0.004(0.004)	-0.001(0.005)
Multiplicity	0.471(0.11)*	0.501(0.119)*
Tumor size (3-5cm)	0.001(0.188)	-0.04(0.189)
Tumor size ($\geq$5cm)	0.229(0.257)	0.226(0.271)
Stage I	-0.121(0.131)	-0.189(0.14)
Stage II	-1.357(0.301)*	-1.51(0.315)*
Grade 2	0.10(0.107)	0.069(0.111)
Grade 3	0.287(0.154)	0.269(0.167)
# of prev.rec	-0.006(0.021)	0.098(0.026)*
θ	0.445	0.719

* significant at 5% level of significance, + Counting process

The frailty model with counting process time scale has higher predictive accuracy for the first recurrence event as compared to the gap time (Table 6). This model without the frailty term is exactly the same as the Andersen-Gill model. Thus, to make the two models comparable, the frailty model with counting process time scale is selected for further interpretation. A variance of zero ($\theta = 0$) would indicate that the frailty component does not contribute to the model. A likelihood ratio test (LRT) for the hypothesis $H_0 : \theta = 0$ was used to assess the significance of the frailty term by fitting a model with and without the frailty terms. The LRT resulted with a chi-square of 42.0 which is tested with a mixture of two chi-squares with zero and one degrees of freedom, yielding a high significant p-value <0.0001. The correlation between recurrent events from the same patient can be estimated by $\rho = \theta/(\theta + 2)$= 0.264. Since the frailty term was significant, in general, the covariate coefficients are not expected to be the same in the Andersen-Gill and frailty models. The two models can result in same parameter estimates when the within recurrent events correlation is zero.

3.4 Risk Prediction Models

From both, the Andersen-Gill and the frailty models, gender, tumor multiplicity, stage and number of previous recurrence were identified as important risk factors for recurrence of bladder cancer. Interpretation in terms of hazard ratio and 95% confidence intervals for the risk factors can be made from the results presented in Table 8.

Table 8: Hazard ratio with 95% CI for AG and frailty models

	Andersen-Gill		Frailty	
	HR	95% CI	HR	95% CI
Male	0.785	(0.649, 0.948)	0.732	(0.563, 0.952)
Age	0.997	(0.99, 1.004)	0.999	(0.99, 1.008)
Multiplicity	1.608	(1.306, 1.98)	1.65	(1.299, 2.097)
Tumor size (3-5cm)	0.877	(0.625, 1.231)	0.961	(0.657, 1.405)
Tumor size ($\geq$5cm)	1.097	(0.667, 1.802)	1.253	(0.704, 2.231)
Stage I	0.794	(0.606, 1.04)	0.828	(0.629, 1.09)
Stage II	0.287	(0.17, 0.485)	0.221	(0.119, 0.409)
Grade 2	1.072	(0.863, 1.332)	1.071	(0.861, 1.332)
Grade 3	1.234	(0.901, 1.69)	1.308	(0.944, 1.813)
# of prev.rec	1.308	(1.265, 1.354)	1.103	(1.048, 1.161)

Based on the Andersen-Gill model, male patients have 22% reduction in risk of having recurrence than female patients. Patients with multiple tumor have higher risk of recurring bladder cancer, 61%, as compared to patients with solitary tumor. Further, compared to patients with stage 0 tumor cancer, patients in the stage II tumor group had 71% reduction in the risk of having the bladder cancer again. The hazard ratio of number of previous recurrence was estimated to be 1.31 with 95% CI of (1.27, 1.35). This means that, for a unit increase in the number of previous recurrence, the risk of recurring the cancer tumor increased by 31% (see Table 8). When using the frailty model, however, reported hazard ratios carry this usual interpretation only if comparing two hazards conditional on a given frailty term u. For example, one would interpret the hazard ratio for the covariate gender as; all other things equal (including the frailty u), the hazard ratio for male is 0.73 times the hazard for females. Comparing two individuals with the same level of frailty and controlling for the other covariates, the risk of having recurrence for a patient with multiple tumor is 1.65 times higher than a patient with solitary tumor. The rest hazard ratios hold similar interpretation. As observed from the estimated hazard ratios from the two models, the frailty model has slightly wider confidence intervals. However, the frailty model has higher discriminatory power than the other marginal models (see Table 6). The cumulative hazard plots in Figure 5(b) during the follow up shows that both models

have the same estimated cumulative hazard in the first five years of the study.

Using the risk factors, one of the important outcomes of the models is patient-specific survival or cumulative risk curves. That is, given the state of a patient's disease at present, the expected future or "natural history" survival for that patient can be predicted. The estimated baseline cumulative risk of bladder tumor recurrence during the follow up for male and female patients aged 65 years with solitary tumor and $< 3cm$ tumor diameter, in stage 0, grade 1 and with no previous recurrence are compared. The plot of the baseline cumulative hazard presented in Figure 5(a) indicate that female patients have the higher risk of having bladder cancer recurrence than male patients throughout the follow up. The estimated survival probability for these male and female patients at one year is 0.81 and 0.76 respectively.

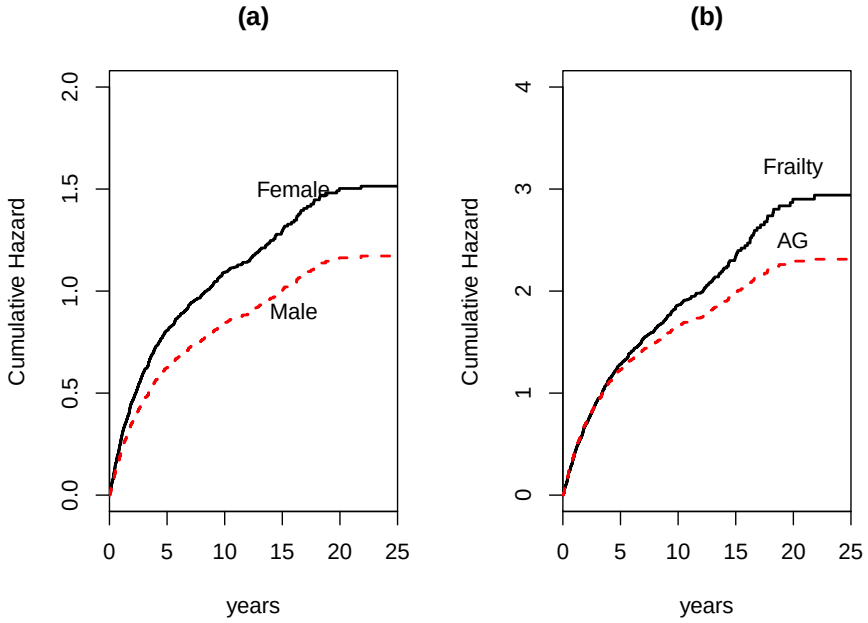

Figure 5: *(a) Baseline cumulative hazard for males and females aged 65 years having solitary tumor, <3cm tumor diameter, in stage 0, grade 1 with no previous recurrence. (b) Cumulative hazard plot for tumor recurrence over the course of the follow-up among 615 patients for AG and Frailty models.*

4 Discussion and Conclusion

A great deal of research is currently underway to develop risk prediction models to estimate the probability that an individual, given risk factors, will develop a specific disease over a defined period of time. Development of these risk prediction models are common practices in clinical studies on either a single or multiple event of interests. Bladder cancer recurrence event is one typical example that a tumor occurs more than once per subject over the follow up time. In this report, 1280 records from 615 bladder cancer patients followed-up during the study period are analyzed. Patients experience 0 to a maximum of 15 recurrence. Observations from the same patient are expected to be correlated, and the ordinary Cox proportional models are not appropriate methods. Different marginal and random-effect approaches which are extensions of the Cox models have been proposed to analyze such data.

In this study, *Andersen-Gill*, *Wei-Lin-Weissfeld* (marginal), *Prentice-Williams-Peterson* (conditional I and II) and gamma shared frailty models were applied to the analysis of recurrence events in bladder cancer patients. In these models, factors such as: gender, tumor multiplicity, stage and number of previous recurrence were identified as important prognostic factors to the recurrence of bladder cancer. No significant difference was found in the effects of these risk factors across the recurrent events. Even though event-specific effect of the covariates was not significant, such analysis enables one to investigate the effect of each risk factors across the strata. However, this approach is inefficient when the number of covariates and events are large, because too many parameter to be estimated.

The major concerns of the different marginal models are the definition of the risk sets and the choice of time scales. In the WLW model, each individual is considered to be at risk of all recurrent events from the start of the observation period while the PWP models assume that an individual is at risk of the k^{th} event only if the person experienced the $(k-1)^{th}$ event. The Andersen-Gill model is different from the PWP and WLW models in its assumption of a common hazard and covariate effect across strata. Fitting all the different marginal models to the analysis of recurrent events are recommended as they provide somewhat different insight. If one is only interested in the overall rate for recurrences of the same nature, the easiest and appropriate method is to use AG model. When the main interest lies on the gap time, conditional II model is the best choice. The WLW model has been criticized because of its failure to accommodate the ordered nature of the recurrent events. Thus, the conditional models will be the appropriate choice when the event order is very important.

All the marginal approaches are based on a "marginal" regression models after correcting

the standard errors and the estimated coefficients carry "population average" interpretations. The common estimates of the covariate effect are unbiased and approximately similar from the AG and the two conditional models, while the parameter estimates and the robust standard errors were inflated (overestimated) in the case of WLW model. However, the conditional II model has the smallest standard errors.

One aim of this study is to propose an appropriate model to predict recurrent rates for new patients, thus, the predictive accuracy and discrimination ability of the models were compared. The Harrell$'s$ C-index (concordance probability) at the first four recurrence events shows that the models have comparable discriminatory ability (ranges from 0.54 to 0.61) except the marginal model, since the parameter estimates from this model are biased.

The last method considered here for analyzing recurrence event times is the random-effect, gamma shared frailty model. It is a natural way of modeling dependent data. The frailty model with counting process time scale is the same as the AG model with a random effect. This random effect induces dependence among the multiple event times. Results of these models show different (larger) parameter estimates and standard errors. Further, a test for the significance of the frailty terms indicates the presence of unobserved variation among patients. Multiple events from same patients share a common frailty. The estimated intra-cluster correlation between observations of a patient is 0.264. Compared to the marginal models, the frailty models have better predictive accuracy and discriminatory ability. It has an estimated concordance probability of 0.84 to discriminate patients with first recurrence into subgroups. The interpretations of the hazard ratio in the marginal and frailty models are quit different in that, the hazard ratio in the frailty models can hold the the usual interpretation within patients sharing the same frailty u_i.

There are some limitations in this study. The models did not include some important patient and tumor characteristics, such as smoking status and number of tumors which might be associated with bladder recurrence, for lack of such information in the dataset. The models predictive ability is assessed or validated internally due to the absence of external dataset or large dataset. Thus, external validation of the models is needed in future investigation.

In conclusion, appropriate approaches to model multiple recurrent events for bladder cancer patients were assessed and compared. Among the marginal models, the Andersen-Gill, and from the random-effect models, the gamma shared frailty model with counting process time scale are identified as the most suitable approaches based on their predictive and discrimination ability. The Andersen-Gill model have marginal-wise interpretation while the frailty model is conditional on patient level.

5 References

1. Freedman, A. N., Seminara, D., Gail, M. H., Hartge, P., Colditz, G. A., Ballard-Barbash, R. and Pfeiffer, R .M. (2005). Cancer risk prediction models: A workshop on development, evaluation, and application. *Journal of the National Cancer Institute*: **97**(10).

2. Cook, R. J. and Lawless, J. F. (2007). The Statistical Analysis of Recurrent Events, Statistics for Biology and Health, Springer.

3. Aalen,O. O., Borgan, O. and Gjessing, H. K. (2008). Survival and Event History Analysis: A Process Point of View. Statistics for Biology and Health, Springer.

4. David, G. Kleinbaum, D. G. and Klein, M. (2005). Survival Analysis: A Self-Learning Text. Second Edition, Springer.

5. Steinberg, G. D. and Katz, M. H. (2011). emedicinhealth, Bladder cancer. Accessed on 7/16/2011. $http://www.emedicinehealth.com/bladder_cancer/article_em.htm$

6. National Cancer Institutes, (2011). Bladder cancer treatment. Accessed on 7/16/2011 $http://www.cancer.gov/cancertopics/pdq/treatment/bladder/Patient/page1$

7. MedScape, (2011). Bladder Cancer, Accessed on 7/16/2011 $http://emedicine.medscape.com/article/438262 - overviewtheCoxModel$. Statistics for Biology and Health. Springer.

8. American Cancer Society, (2011). Learn about Cancer: How bladder cancer is staged? Accessed on 7/16/2011. $http://www.cancer.org/cancer/bladdercancer/detailedguide/bladder-cancer-staging$

9. MedicaLook, Your Medical World: Bladder Cancer (n/d). Accessed on 7/16/2011 $http://www.medicalook.com/Cancer/Bladder_cancer.html$

10. Cancer HelpUK, (2010). Bladder cancer stage and grade. Accessed on 7/16/2011. $http://cancerhelp.cancerresearchuk.org/type/bladder-cancer/treatment/bladder-cancer-stage-and-grade$

11. Therneau, T. M. and Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health. Springer.

12. Castaneda, J. and Gerritse, B. (2010). Appraisal of several methods to model time to multiple events per subject: Modeling time to hospitalizations and death. *Revista Colombiana de Estadstica*, **33**: 43-61.

13. Wienke, A. (2011). Frailty Models in Survival Analysis. Chapman & Hall CRC

14. SAS/STAT(R) 9.2 User's Guide, Second Edition

15. Hosmer, D. W. and Lemeshow, S. (1999). Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley Series in Probability and Statistics.

16. Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study, *Annals of Statistics* , **10**(4): 1100-1120.

17. Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association*, **84**: 1065-1073.

18. Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data, *Biometrika*, **68**: 373-379.

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

20. Johnson, C. J., Boyce, M. S., Schwartz, C. C. and Haroldson, M. A. (2004). Modeling survival: Application of the Andersen-Gill model to yellowstone grizzly bears, *The Journal of Wildlife Management*; **68**: 966-978.

21. Wei, L. J. and Glidden, D. V. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in medicine*, **16**: 833-839.

22. Metcalfe, C. and Thompson, S. G. (2007). Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: Should it be applied to a recurrent events outcome? *Statistical Methods in Medical Research*; **16**: 103-122

23. Jisheng Cui, J., Forbes, A., Kirby, A., Marschner, I., Simes, J., West, M. and Tonkin, A. (2008). Parametric conditional frailty models for recurrent cardiovascular events in the lipid study. Clinical Trials; **5**: 565-574.

24. Box-Steffensmeier, J. M. and Boef, S. D. (2006). Repeated events survival models: The conditional frailty model. *Statistics in medicine*, **25**: 3518-3533.

25. Therneau, T. M., Grambsch, P. M. and Pankratz, V. S. (2000). Penalized survival models and frailty. Technical report #**66**.

26. Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R. and Sylvester, R. (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics and Data Analysis*, **40**: 603-620.

27. Duchateau, L., Janssen, P., Kezic, I. and Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Appl. Statist.* **52**(3): 355-363

28. Duchateau, L., Nguti, R. and Janssen, P. (n/d). Report on frailty model.

29. Burton, A. and Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guideline. *British Journal of Cancer* **91**: 4-8.

30. Marshall, A., Altman, D. G., Royston, P. and Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modeling studies: A simulation study. *BMC Medical Research Methodology*,**10**: 7

31. Marshall, A., Altman, D. G., and Holder, R. L. (2010). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Medical Research Methodology*,**10**: 112

32. Greenland, S. and Finkle W. D. (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology*, **142** (12): 1255-64.

33. Berglund, P. A. (2010). An introduction to multiple imputation of complex sample data using SASv9.2. SAS Global Forum.

34. Rogier, A., Dondersa, T., van der Heijdenc, G. J., Stijnend, T. and Moonsc, K. G. (2006). Review: A gentle introduction to imputation of missing values. Special Series: Missing data. *Journal of Clinical Epidemiology* **59**: 1087-1091

35. Lusa, L., Miceli, R. and Mariani, L. (2007). Estimation of predictive accuracy in survival analysis using R and S-plus. *Computer methods and programs in biomedicine*. **87**: 132-137.

36. Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. and Kattan, M. W., (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, **21**(1).

37. Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics*, **56**: 249-255.

38. Taktak, A. F., Eleuteri, A., Lake, S. P. and Fisher, A. C. (2002). Evaluation of prognostic models: Discrimination and calibration performance.

39. Harrell, Jr., F. E., Lee, K. L. and Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy and measuring and reducing errors. *Statistics in medicine*, **15**: 361-387.

40. Pencina, M. J. and D'Agostino, R. B. (2004). Overall C as a measure of discrimination in survival analysis model specific population value and confidence interval estimation. *Statistics in medicine*, **23**: 2109-2123.

41. Shariat, S. F., Margulis, V., Lotan, Y., Montorsi, F. and Karakiewicz, P. I. (2008). Nomograms for bladder cancer. *European urology*, **54**: 41-53.

42. Harrell, Jr. F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, **247**(18).

43. Oirbeek, R. V and Lesaffre, E. (2009). A new discrimination index for frailty survival models.

44. Gonen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, **92**(4): 965 -970

6 Appendix

Plots of scholfeld residuals based on Andersen-Gill model for checking the proportionality hazard assumptions.

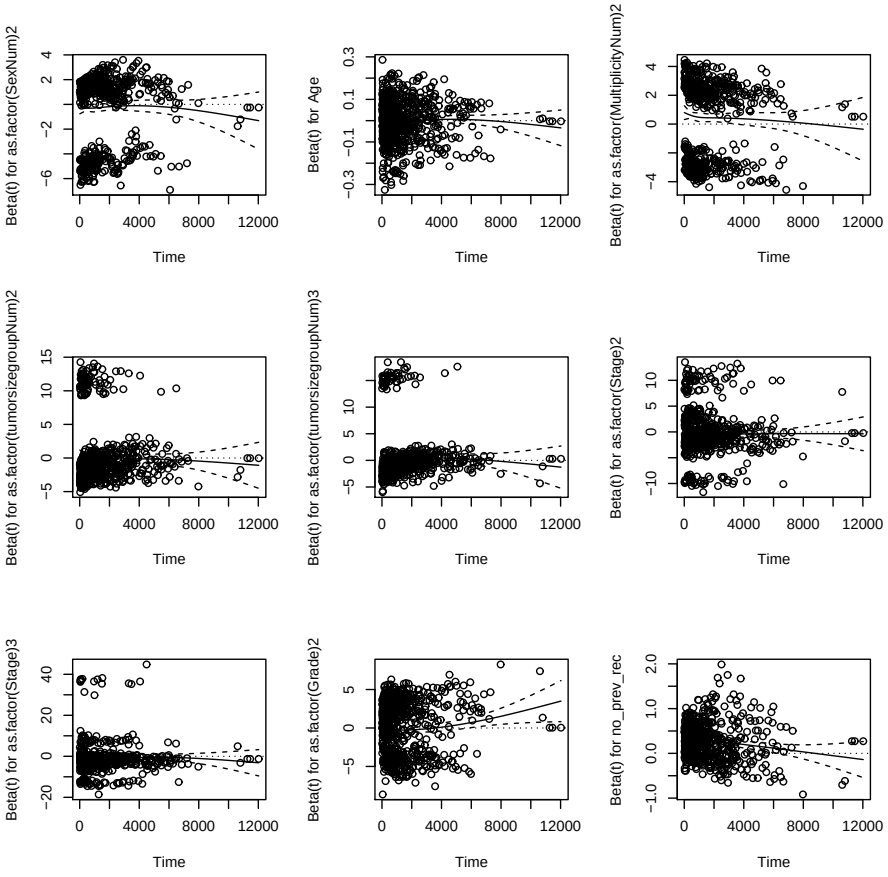

Figure 6: Schoenfeld residuals for some selected variables with ± 2 standard errors

Some R codes used to fit the models

```
library(foreign)
library(survival)
library(rms)
library(CPE)
library(Design)
library(mice)
library(survcomp)
```

```
data1<- read.spss("D:/thesis work/trial/missing.sav", to.data.frame=TRUE)
names(data1)
Agem<-data1$Age-65
##### Multiple imputation####
data2<- mice(data1,m = 5,seed=1000,Method =c(
"","","","","","","","","","",
"","","polyreg","logreg","logreg","",""))
data3=complete(data2,1)
data4<-complete(data2,2)
data5<-complete(data2,3)
data6<-complete(data2,4)
data7<-complete(data2,5)
#### 1. AG model
AG1<-with(data2,coxph(Surv(Tstart, Tstop, event)~ as.factor(SexNum)+
      Agem+ as.factor(MultiplicityNum)+as.factor(tumorsizegroupNum)+
      as.factor(Stage)+ as.factor(Grade)+  cluster(Pat_ID), x=T))
AG2<-summary(pool(AG1))
AG3<-round(summary(pool(AG1)),4)
#### 2. Conditional I model
CI1<-with(data2,coxph(Surv(Tstart, Tstop, event)~ as.factor(SexNum)+
     Agem+ as.factor(MultiplicityNum)+as.factor(tumorsizegroupNum)+
     as.factor(Stage)+ as.factor(Grade) + cluster(Pat_ID)+
     strata(Sequence), x=T, y=T))
CI2<-summary(pool(CI1))
CI3<-round(summary(pool(CI1)),4)
#### 3. Conditional II model
CII1<-with(data2,coxph(Surv(Time, event)~ as.factor(SexNum)+ Agem+
       as.factor(MultiplicityNum)+as.factor(tumorsizegroupNum)+
       as.factor(Stage)+ as.factor(Grade) + cluster(Pat_ID)+
       strata(Sequence), x=T, y=T))
CII2<-summary(pool(CII1))
CII3<-round(summary(pool(CII1)),4)
#4. Marginal WLW model with new dataset
M1<-with(datamm2,coxph(Surv(Time_nth_rec, event)~ as.factor(SexNum)+
    Agem+as.factor(MultiplicityNum)+as.factor(tumorsizegroupNum)+
    as.factor(Stage)+ as.factor(Grade) + cluster(Pat_ID)+
    strata(Sequence), x=T, y=T))
```

```
M2<-summary(pool(M1))
M3<-round(summary(pool(M1)),4)
### 5. frailty model with gap time
frailty1<-with(data2,coxph(Surv(Time, event)~ as.factor(SexNum)+ Agem+
     as.factor(MultiplicityNum)+as.factor(tumorsizegroupNum)+
     as.factor(Stage)+ as.factor(Grade) + no_prev_rec+
     frailty(Pat_ID, dist="gamma", sparse=FALSE), x=T, y=T))
frailty2<-summary(pool(frailty1))
frailty3<-round(summary(pool(frailty1)),4)
### 6. frailty model with CP time scale
frailty.CP<-with(data2,coxph(Surv(Tstart, Tstop, event)~ as.factor(SexNum)+
        Agem+as.factor(MultiplicityNum)+as.factor(tumorsizegroupNum)+
        as.factor(Stage)+ as.factor(Grade) +no_prev_rec+
        frailty(Pat_ID, dist="gamma", sparse=FALSE), x=T, y=T))
frailty.CP2<-summary(pool(frailty.CP))
frailty.CP3<-round(summary(pool(frailty.CP)),4)

#### sample code for computing the C-statistics for AG model
ag1<-coxph(Surv(Tstart, Tstop, event)~ as.factor(SexNum)+ Agem+
        as.factor(MultiplicityNum)+as.factor(tumorsizegroupNum)+
        as.factor(Stage)+ as.factor(Grade)+ no_prev_rec +
        cluster(Pat_ID), x=T,y=T, data=data3)
AG.X<-ag1$x[data3$Sequence==1,]
pi.AG<-AG.X%*%ag1$coef
surv<-Surv(data6$Time, data3$event)[data3$Sequence==1,]
c=rcorr.cens(pi.AG, surv)  #1st recurrence
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
The frailty model versus the Andersen-Gill model for the prediction of recurrent events

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Baye, Fitsum Megersa

Datum: **12/09/2011**