Journal of Statistical Research *ISSN 0256 - 422 X*
2017, Vol. 51, No. 2, pp. 145-164

SURVIVAL ANALYSIS OF RECURRENT EVENTS ON PROSTATE CANCER: FACTS FROM CANCER GENOME

Farhin Rahman
Ball State University, Muncie, Indiana 47306, USA
Email: frrahman@bsu.edu

Munni Begum*
Ball State University, Muncie, Indiana 47306, USA
Email: mbegum@bsu.edu

SUMMARY

Many diseases and clinical outcomes may recur to the same patient. These events are termed as recurrent events. Several statistical models have been proposed in the literature to analyze recurrent events. In this study, we identify the clinical and the genetic risk factors for recurring tumors among prostate cancer patients from The Cancer Genome Atlas (TCGA). Five statistical approaches for modeling recurrent time-to-event are implemented to identify and to determine the effects of the clinical and the genetic risk factors of tumor recurrence. In particular, we consider Andersen-Gill (A-G), Wei-Lin-Weissfeld (WLW), Prentice-Williams-Peterson Total Time (PWP-TT), Prentice-Williams-Peterson Gap Time (PWP-GT) and Frailty models. We present and discuss the risk factors influencing the recurrence of tumors and their impacts in prostate cancer patients obtained from five commonly used models in this paper.

Keywords and phrases: Survival Modelling, Recurrent Events, Gene Expression, Prostate Cancer, TCGA

1 Introduction

Many applications involve repeated events where a subject may experience a single event repeatedly or multiple events over the life time. For example, in cancer studies, it is of interest to identify risk factors and therapeutic measures to multiple occurrences of a cancer. The event which occurs more than once to a participant over the follow-up time is often termed as recurrent event, such as, admissions to the hospitals, repeated heart attacks for coronary patients, recurrence of tumors in cancer patients. Recurrent events provide more information about the disease progression than a single event. When the outcome variable of interest in survival analysis is a recurrent event, single event models such as, Cox proportional hazards model or parametric accelerated failure time models do not provide

* Corresponding author

optimal results. In this paper we study, the prostate cancer survival data obtained from the cancer genome atlas (TCGA) (The Cancer Genome Atlas: www.gdc.nci.nih.gov). Both clinical and genetic risk factors (RNASeq expressions of a number of genes associated with prostate cancer) are available in this data. A little over fifteen percent of the patients with prostate cancer experienced recurrence of tumors.

The primary objective of this paper is to identify the clinical risk factors of recurrence of tumors in prostate cancer patients in the presence of genetic information in form of gene expression measures. We also compare the most commonly used statistical methods for analyzing recurrent time to event data, including Andersen-Gill (A-G), Wei-Lin-Weissfeld (WLW) marginal mean/rate model, Prentice-Williams-Peterson Total Time (PWP-TT), Prentice-Williams-Peterson Gap Time (PWP-GT) and Frailty model. A number of studies investigated clinical (Burns et al. 2014; Chan, 2013; Kenfield et al. 2011; Ross et al. 2003; Wilding et al. 2014) and genetic factors (Barbieri et al. 2012; Jhun et al. 2017; Mortensen et al. 2015; Robinson et al. 2015; Shancheng et al. 2012; William et al. 2014) separately. Our goal is to determine which clinical and genetic factors are jointly responsible on the recurrence of tumors among prostate cancer patients so that appropriate measure or treatment can be implemented to avoid tumor recurrences. In addition we compare these results obtained from the five models for analyzing recurrent time to event in practice.

2 Background

Prostate cancer is one of the most common types of cancer in men. American Cancer Society (American Cancer Society: www.cancer.org) has declared prostate cancer as the third leading cause of cancer death in American men, behind lung cancer and colorectal cancer. According to National Cancer Institute (NCI) (National Cancer Institute: www.cancer.gov), there were 161,360 new cases of prostate cancer and an estimated 26,730 people might die of this disease in 2017. Some risk factors that may have clear effect on prostate cancer are age, race, geography, family history, and changes in genes whereas some may have less effect on prostate cancer such as, diet, obesity, smoking status, chemical exposures, sexually transmitted infections, inflammation of the prostate (American Cancer Society: www.cancer.org).

Several studies have been implemented to identify clinical factors which might be responsible for developing prostate cancer. One study identified the existence of racial/ethnic disparities by comparing the average mean tumor size, the median of survival time, and the survival function between White and African American men (Chan, 2013). On the other hand, to examine the role of health care provider and socio-economic status (area-based deprivation) on survival, age, stage, Gleason grade, marital status and region of residence are considered as the controlling factors for prostate cancer survival (Burns et al. 2014). Another study revealed that stage, grade of tumor, level of prostate-specific antigen (PSA) along with race were found as the strongest prognostic factors in period analysis of prostate cancer survival (Wilding et al. 2005). Smoking status was also taken in to ac-

count for prostate cancer survival and recurrence along with the presence of cardiovascular disease (Kenfield et al. 2011) in one study. To identify the correlation of primary tumor prostate-specific membrane antigen expression with disease recurrence in prostate cancer, tumor grade, pathological stage, Gleason grade, level of PSA, biochemical recurrence were considered primarily (Ross et al. 2003).

In addition to clinical factors, genetic factors play significant role in the development of cancer, including prostate cancer. A large number of studies on identifying responsible genes on the development of prostate cancer have been carried out. Some studies only focus on the significant genes and their genomic expression in finding the impact on the survival of prostate cancer patients. Several genes and chromosomal regions have been found to be associated with prostate cancer in various linkage analyses, case-control studies, genome-wide association studies (GWAS), and admixture mapping studies. One such study was conducted by the health professionals of the National Cancer Institute (NCI) (National Cancer Institute: www.cancer.gov) in order to better understand and address psychosocial and behavioral issues in high-risk families. Another study on prostate cancer tissue found twelve genes as independent predictors of recurrence after prostatectomy (Jhun et al. 2017). A huge number of genes which are responsible for prostate cancer were found from different types of studies, such as, therapeutic potential on the recurrence rearrangements in prostate cancer patients (Mortensen et al. 2015), recurrent copy number alterations in prostate cancer (William et al. 2014), RNA Sequencing analysis (Shancheng et al. 2012), mutation (Barbieri and Baca, 2012), Integrative Clinical Genomics (Robinson et al. 2015) and so on.

Different recurrent time to event analysis were proposed over time. The most frequent methods in analyzing recurrent events were used for different studies are Andersen-Gill model (A-G)- an extension of Cox proportional hazards model which assumes that the correlation between event times for a person can be explained by past events (Amorim and Cai, 2015; Kelly and Lim, 2000; Ullah et al. 2012), Prentice-Williams-Peterson Total Time (PWP-TT) model- evaluates the effects of a factor for a event since the entry time in the study (Amorim and Cai, 2015; Kelly and Lim, 2000; Ullah et al. 2012), Prentice-Williams-Peterson Gap Time (PWP-GT) model- evaluates the effect of a factor for a event since the time from the previous event (Amorim and Cai, 2015; Kelly and Lim, 2000; Ullah et al. 2012), Wei-Lin-Weissfeld (WLW) model- deals with cumulative time from randomization to events (Amorim and Cai, 2015; Kelly and Lim, 2000; Ullah et al. 2012), Frailty model- induces dependence among the multiple event times (Amorim and Cai, 2015; Oakes, 1992; Ullah et al. 2012), Multi-State model- provides a means of analyzing data with multiple event times (Amorim and Cai, 2015), Accelerated Failure Time model- regresses the logarithm of the survival time over the factors (Wei and Glidden, 1997; Zare et al. 2015), and so on. Every model for analyzing the effects of recurrent events on the survival differs in assumptions and interpretation. Also, they have advantages as well as disadvantages. However, researchers need to be careful in selecting appropriate models based on the research questions, data structure and information given about subjects, etc.

3 Methodology

In time to event analysis, we usually refer to the time variable as survival time or time to event, because it is the time that an individual has 'survived' over some follow-up period. We also typically refer to the event as a failure, because the event of interest usually is death, disease incidence, or some other negative individual experience. In this study time to event refers to the 'time to tumors after initial treatment'. For some patients, tumor recurred more than once. Thus our analysis calls for methods to analyze recurrent events where the events or failures are the recurrence of tumors.

Since the main objective of this study is to identify the important clinical and genetic risk factors that may prevent recurrences, we employ several commonly used methodologies to analyze recurrent tumor events that occurred to the prostate cancer patients under considerations and compare the results thereof. These methods are discussed briefly as follows.

3.1 The Andersen and Gill Model

The Andersen and Gill (A-G) is based on the counting process approach and generalizes the Cox model (Andersen and Gill, 1982). The A-G model assumes that the recurrent events are identical and is formulated in terms of increments in the number of events along the time line. In other words, the A-G model is formulated in terms of relating the intensity function for the kth recurrent event of the ith subject to the covariates as follows.

$$\lambda_{ik}(t \mid \mathbf{Z}_{ik}(t)) = Y_{ik}(t)\lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_{ik}(t)), \quad k = 1, 2, \ldots, K.$$

Here, $\lambda_{ik}(t)$ represents the hazard or intensity function for the kth recurrent event of the ith subject at time t; $\lambda_0(t)$ represents the common baseline intensity function for all events over time; $\mathbf{Z}_{ik}(t) = (Z_{1ik}(t), Z_{2ik}(t), \ldots, Z_{pik}(t))$ is the vector of possibly time-dependent covariates; $\boldsymbol{\beta}$ is a fixed vector of p coefficients; and $Y_{ik}(t)$ is a predictable process taking values 1, when ith subject is under observation, and 0, otherwise. A-G models considers the outcome of interest as time since randomization for a treatment (or other exposure) until an event occurs. Thus the outcome is time since the study entry, also known as total time scale. A common baseline hazard function and the same regression coefficients for all the events are considered.

A-G model assumes that the correlation between event times for a person can be explained by past events, which implies that the time increments between events are conditionally uncorrelated, given the covariates. It is a suitable model when correlations among events for each individual are induced by measured covariates. Thus, dependence is captured by appropriate specification of time-dependent covariates, such as number of previous events or some function thereof. However, if this assumption does not hold, robust sandwich covariance matrix for the resulting regression coefficient estimators are calculated that address the correlations among the recurrent event times.

3.2 Prentice, Williams, and Peterson Model

Prentice, Williams and Peterson (PWP) (Prentice, Williams and Peterson, 1981) proposed the first extended Cox models for multiple events. More specifically they proposed regression models for the intensity or hazard function for kth recurrent event of ith subject on two different time scales, namely, the total time (PWP-TT) and the gap time (PWP-GT). PWP-TT and PWP-GT models for the intensity or hazard function for the kth recurrent event of ith subject can be written as follows.

$$\lambda_{ik}(t \mid \mathcal{N}_i(t), \mathbf{Z}_{ik}(t)) = Y_{ik}(t)\lambda_{0k}(t)\exp(\boldsymbol{\beta}_k'\mathbf{Z}_{ik}(t)), \quad k = 1, 2, \ldots, K.$$

$$\lambda_{ik}(t \mid \mathcal{N}_i(t), \mathbf{Z}_{ik}(t)) = Y_{ik}(t)\lambda_{0k}(t - T_{i,k-1})\exp(\boldsymbol{\beta}_k'\mathbf{Z}_{ik}(t)), \quad k = 1, 2, \ldots, K.$$

It is to be noted that the intensity functions in both models depend on the event history process $\mathcal{N}_i(t)$ for the ith subject in addition to the covariate process $\mathbf{Z}_{ik}(t)$. In the first model (PWP-TT), the time scale is total time t from the beginning of the study and in the second model (PWP-GT), the time scale is the time $t - T_{i,k-1}$, from the immediately preceding failure. One important distinction between PWP models and A-G model is that the baseline intensity functions $\lambda_{0k}(t)$ and the regression coefficients $\boldsymbol{\beta}_k$ for PWP models are event specific, whereas the baseline intensity function, $\lambda_0(t)$ and the regression coefficients are the same for A-G model across all recurrent events. Similar to A-G model, $Y_{ik}(t)$ is defined as a predictable process taking values, 1 when ith subject is under observation, and 0, otherwise.

3.3 Wei, Lin, and Weissfeld Marginal Model

Marginal means models are proposed by Wei, Lin, and Weissfeld (WLW) (Wei, Lin and Weissfeld, 1989) to analyze multivariate failure times in a regression setup. This general approach can be adopted to analyze recurrent events with assumption that there is no time dependent covariates or there is no specific dependence structures among the recurrent event times within a subject. Under the marginal means model, the intensity or hazard function for the kth recurrent event of ith subject has the following form.

$$\lambda_{ik}(t) = \lambda_{0k}(t)\exp(\boldsymbol{\beta}_k'\mathbf{Z}_{ik}(t)), \quad k = 1, 2, \ldots, K.$$

Similar to PWP models marginal mean models assumes event specific baseline intensity function $\lambda_{0k}(t)$ as well as event specific regression coefficients $\boldsymbol{\beta}_k$. Among others, one major issue in applying WLW marginal modeling approach is that it allows a subject to be at risk for several events simultaneously. It is to be noted that this approach considers all recurrent events occurred to the same subject as a single counting process and does not consider the subject's past event history.

3.4 Frailty Model

Frailty models also known as random effect models are in which the intensity or the hazard for the event of interest depends partly on an unobservable random variable and is assumed

to act multiplicatively on the intensity (Liu et al. 2004; Nielsen et al. 1992). Within subject correlations are modeled explicitly in frailty models. For multiple event scenario, an unobservable random variable induces dependence among the these event times of the same subject. Random effects or frailty models can be applied to model the conditional intensity or hazard function of recurrent events, when there is heterogeneous susceptibility to the risk of such events (Amorim and Cai, 2015). Under frailty or random effects model, the intensity or hazard function for the kth recurrent event of ith subject has the following form.

$$\lambda_{ik}(t \mid W_i) = W_i \lambda_0(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_{ik}(t)), \;\; k = 1, 2, \ldots, K.$$

Here, the frailty term $W_i, i = 1, \ldots, n$, are assumed to be independent and identically distributed with a common parametric density. Known as shared or Gamma frailty model (Clayton and Cuzick, 1985), where W_i has a Gamma distribution with mean 1 and some variance θ.

4 Data and Variables

Prostate cancer survival data is obtained from the cancer genome atlas (TCGA) (The Cancer Genome Atlas: gdc.nci.nih.gov), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). It has comprehensive and multi-dimensional maps of the key genomic changes in 33 types of cancer. TCGA created a genomic data analysis pipeline that can effectively collect, select, and analyze human tissues for genomic alterations on a very large scale. This database consists of information separately on clinical characteristics, genome, DNA sequencing, mutation, methylation, miRSeq and mRNASeq. In this study, both clinical and RNAseq gene expressions data set are accumulated to identify the clinical and genetic risk factors for prostate cancer patients. The prostate cancer data has follow-up periods of four years (2011 to 2015) with cancer initialization year of 2000.

4.1 Response Variable

Recurrence Time to Tumors

We consider recurrent time to tumors after initial treatment as the response or outcome variable. Time to tumors indicates the number of days an individual has developed tumors. Three hundred and ninety seven patients are included in the model while considering recurrent events after omitting missing observations. We considered the recurrence of tumor in different follow-up times. A total of 61 patients are detected to have recurred tumors after initial treatment. Since the data set had follow-up up to second round, in addition to single event, we counted new tumor at 2^{nd} follow-up as additional recurrence. Figure 1 shows the overall tumor free survival for recurrent events.

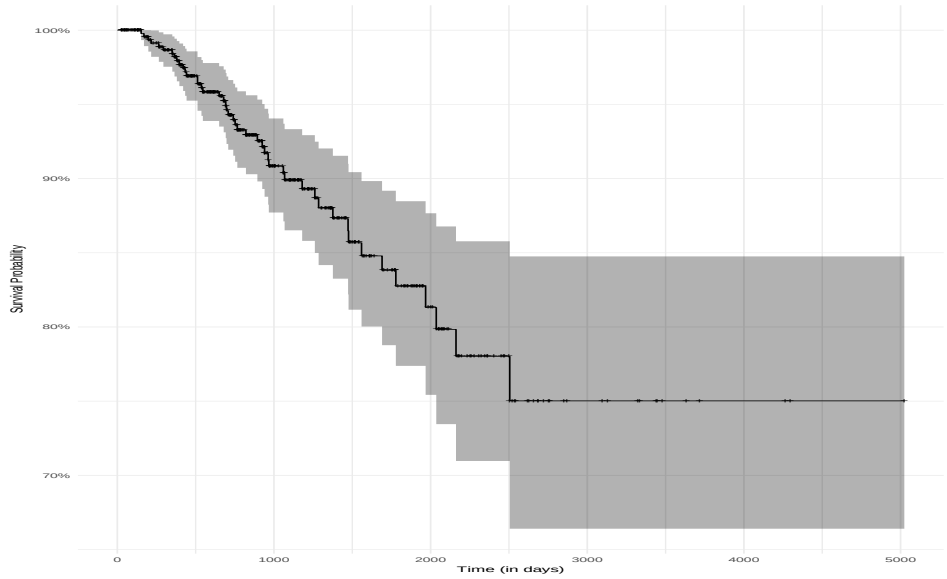

Figure 1: Tumor Free Survival Function for Recurrent Events

4.2 Risk Factors

Clinical Data and Variables

TCGA stores clinical as well as demographic information about the patients along with other genomic information. In prostate cancer data, 337 clinical information on 498 prostate cancer patients are observed. This data consists of information about patients' demography (i.e., age, race, ethnicity, laterality, death reason, etc.), histology, diagnostics, drug, therapy, radiation, tumor as well as follow-up on several health conditions. To determine the factors associated for developing tumors after initial treatment, we finally consider four clinical variables (i.e., age, clinical stage, pathological stage, treatment). These clinical variables are comparable with the risk factors of prostate cancer literature (Burns et al. 2014; Chan, 2013; Kenfield et al. 2011; Ross et al. 2003; Wilding et al. 2014).

Age is considered as one of the responsible factors for tumor growth in cancer patients. Patients diagnosed with prostate cancer in TCGA database has a mean age of 61 years with minimum 41 and maximum 78 years of age. Race is another important factor (Chan, 2013). However, patients' race is not considered in this study. Among different races, the likelihood of developing prostate cancer varies. Information about 147 white, 7 African American, 2 Asian men are enlisted in the clinical data set with missing racial information of 342 patients. High number of missing values and the contradiction of previous study findings that prostate cancer is more likely to occur in African American men lead us to drop race from further considerations.

In choosing treatment options as well as predicting a man's outlook for survival, clinical

and pathological stages are very important. Stage variable helps to identify how far a cancer has spread. Clinical stage is an important predictor to determine the extent of the disease, based on the results of physical exam, lab tests, prostate biopsy and imaging tests, if any. On the other hand, pathological stage is measured based on the results above, plus the results of surgery. This means that if a patient has surgery, the stage of his cancer might actually change afterward. We use the American Joint Committee on Cancer (AJCC) (American Joint Committee on Cancer: www.cancerstaging.org) Tumor-Nodes-Metastasis (TNM) system to create the clinical as well as pathological stages grouping. According to AJCC, both clinical and pathological stages have five categories (I, IIA, IIB, III, IV). For the simplicity as well as to avoid misleading results due to the missing information, we divide the distribution of tumor type in low-stage (I/IIA/IIB) and high-stage (III/IV). Distribution of patients in different stages are shown in figure 2.

Another clinical factor- 'treatment' (whether the patient received any of the treatments - molecular therapy, neoadjuvant therapy, chemotherapy and radiation therapy) is considered based on the treatment the patients received after diagnosis. Among a total of 498 patients only 97 (19%) received treatment (either received molecular, neoadjuvant, chemotherapy or radiation therapy).

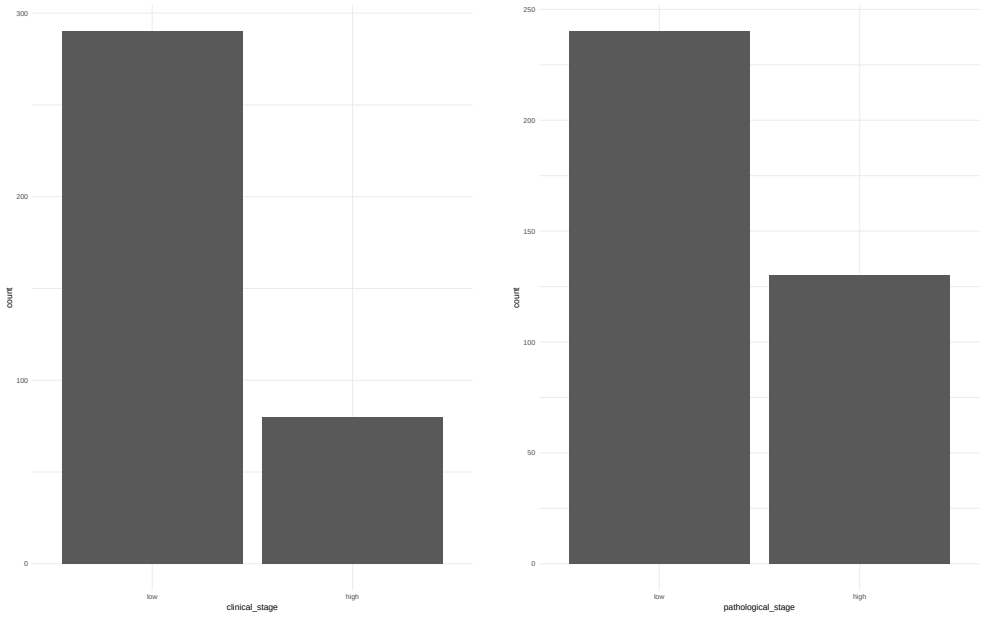

Figure 2: Distribution of Clinical and Pathological Stages

RNA Sequencing Data and Gene Expressions

Deoxyribonuclic acid (DNA) and Ribonuclic acid (RNA) play a fundamental role in carrying genetic information in all living organisms on the earth. RNA directly codes for amino acids and acts as a messenger between DNA and ribosomes to make proteins while DNA stores genetic information. The RNA Sequencing (RNAseq), also termed as whole transcriptome shotgun sequencing is a recent advancement in biological research. The high-throughput sequencing technology sequence complementary deoxyribonucleic acid (cDNA) to extract information on RNA content from the sample and generates millions of short read sequences (deoxyribonucleic acid (DNA) sequences). These short reads are then mapped or aligned to a reference genome. Thus, the number of mapped reads to a gene is used as a measure of tag abundance, or equivalently known as gene expression. Gene expression is quantified by summing the number of sequencing reads that map to exons within each gene. RNAseq methodology was developed in order to comprehensively measured different types of RNA and to quantify their expression levels during various developmental stages and across experimental conditions (Paul, 2010).

In TCGA database of RNAseq, gene expressions of 20531 genes from 550 prostate cancer patients are enlisted. z-scores are calculated for each gene of each patient. For expression measures from RNA-Seq experiments, the standard rule is to compute the relative expression of an individual gene and tumor to the gene's expression distribution in a reference population (The Cancer Genome Atlas: gdc.nci.nih.gov). That reference population is either all tumors that are diploid (containing two complete sets of chromosomes, one from each parent) for the gene in question or when available, normal adjacent tissue. The returned value indicates the number of standard deviations away from the mean of expression in the reference population (z-score). This measure is useful to determine whether a gene is up- or down-regulated relative to the normal samples or all other tumor samples. In this study, we considered those genes with $z > \pm 1.96$ (roughly $p = 0.05$ or 2 SD away) to be differentially expressed. To obtain z-scores for the RNA Seq data we use the following formula:

$$z = \frac{\text{Expression for gene } X \text{ in tumor } Y - \text{Mean expression for gene } X \text{ in normal}}{\text{Standard deviation of expression for gene } X \text{ in normal}}.$$

We used the z-score values to define which samples are altered and which do not change. The numbers for *altered* and *not altered* refer to the number of samples with gene expression higher/lower than a specific threshold such as z-score of 2.

Significant Gene Selection

We measure z-scores for the expression measures of 20531 genes which are available from the prostate cancer samples in TCGA. Top ranked genes (i.e., P-value ≤ 0.10) for which survival are significantly different in altered and non-altered groups are selected using each of the five methods (A-G, WLW, PWP-TT, PWP-GT and Frailty) and are listed in Table 1. If the absolute value of z-score is greater than $1.96 \equiv 2$, i.e., the gene whose z-score lies apart from the 95% of the distribution is considered as 'altered gene' otherwise 'non-altered'.

Table 1: Significant Genes from Recurrent Events

A-G Model	CTHRC1^a	BUB1^a	AURKA^b	PAGE4^b	FAM49B^b	CCNA2^b
	EZH2^c	CDKN3^c	MKI67^c	MAPK3^c	CTGF^c	TOP2A^c
	NOX4^c	GRB2^c	CCL2^c	ALOX15^c	TPX2^c	ELAC2^c
	BIRC5^d	CXCR4^d	SNHG8^d	IL1B^d	CCNB1^d	
WLW Model	CTHRC1^a	FAM49B^a	AURKA^b	PAGE4^b	CCNA2^b	BUB1^b
	GRB2^b	MKI67^b	CDKN3^b	EZH2^b	MAPK3^c	CTGF^c
	CCL2^c	TOP2A^c	NOX4^c	TPX2^c	ALOX15^c	CXCR4^c
	BIRC5^c	ELAC2^c	IL1B^d	CCNB1^d	SNHG8^d	
PWP-TT Model	CTHRC1^a	BUB1^b	FAM49B^b	PAGE4^b	AURKA^b	CCNA2^b
	EZH2^c	GRB2^c	MAPK3^c	MKI67^c	CDKN3^c	TOP2A^c
	CTGF^c	NOX4^c	CCL2^c	ALOX15^c	TPX2^c	CXCR4^c
	BIRC5^c	ELAC2^c	IL1B^d	SNHG8^d	CCNB1^d	PCA3^d
PWP-GT Model	CTHRC1^a	BUB1^b	PAGE4^b	HOXD13^c	FASN^c	TTTY15^c
	NOX4^c	SNHG8^c	CCNA2^c	AKT1^c	MKI67^c	AURKA^c
	TPX2^c	CCL2^c	SDHC^d	EZH2^d	FAM49B^d	TOP2A^d
	ELAC2^d	ALOX15^d	ARAF^d			
Frailty Model	CTHRC1^a	BUB1^b	AURKA^b	PAGE4^b	CCNA2^c	EZH2^c
	FAM49B^c	MAPK3^c	CDKN3^c	MKI67^c	TOP2A^c	CCL2^c
	GRB2^c	NOX4^c	ALOX15^d	TPX2^d	CTGF^d	BIRC5^d

a : P-value $\leq .001$, b : $.001 <$ P-value $\leq .01$, c : $.01 <$ P-value $\leq .05$, d : $.05 <$ P-value $\leq .1$

5 Analysis and Findings

5.1 Andersen Gill (A-G) Model

One of the simplest method to implement recurrent event analysis follows the counting process approach of Andersen-Gill (A-G) (Andersen and Gill, 1982) model and measures the intensity rate. At first, we fit the A-G model on clinical variables and altered status of twenty three genes separately. Apart from clinical stage and treatment status, patient's age (HR: 1.074, P-value: 0.007) along with patients' belong to high (stage III/IV) pathological stage (HR: 1.080, P-value: 0.065) are found significantly associated with the recurrence of tumors. Four genes (SNHG8, CCL2, CTHRC1 and PAGE4) are found to have significant effect on the recurrence of tumors while fitting gene expressions separately.

In contrast, while considering both clinical and gene expressions in the same model we get different results. Outcome of the clinical factors and gene expressions from the combined model are presented in Table 2. Patient's age becomes statistically insignificant. However, hazard rate of 1.041 indicates the expected hazard, i.e., recurrence of tumors is 1.041 times higher for a patient who is one year older. Patients who belong to high (stage III/IV) clinical stage are 66.3% less likely to have recurrent tumors than patients who are at low

Table 2: Effects of Risk Factors: A-G Model

Risk Factors	Hazard Ratio (95% CI)
Age	1.041 (0.941, 1.152)
Clinical Stage (High)	0.337 (0.051, 2.242)
Pathological Stage (High)	4.322^{d} (0.839, 22.259)
Treatment (Not Received)	1.884 (0.186, 19.062)
BUB1 (Altered)	15.333^{c} (1.784, 31.736)
PAGE4 (Altered)	3.602^{d} (0.885, 14.662)
FAM49B (Altered)	5.961^{d} (1.146, 30.982)
MKI67 (Altered)	7.912^{d} (0.830, 17.399)
CTGF (Altered)	3.475^{d} (0.790, 15.279)
CCL2 (Altered)	5.796^{c} (1.359, 24.704)

a : P-value $\leq .001$, b : $.001 <$ P-value $\leq .01$,

c : $.01 <$ P-value $\leq .05$, d : $.05 <$ P-value $\leq .1$

clinical stage. It is also to be noted that patients who belong to high pathological stage have higher risk of having recurring tumors than those at low pathological stage. One of the crucial risk factors of the prostate cancer patients is the treatment. The results from A-G model show that the patients belonging to non-treatment group (did not receive any of the molecular, neoadjuvant, chemotherapy or radiation therapy) have higher relative risk (88.4%) of having recurring tumor than those who received any of the treatments stated above. In addition, patients with altered genes MKI67, BUB1, CTGF, CCL2, FAM49B and PAGE4 have higher risk of developing tumors than those do not have these genes altered.

5.2 Wei-Lin-Weissfeld (WLW) Model

The second model we consider is Wei-Lin-Weissfeld (WLW) model (Wei, Lin and Weissfeld, 1989) that is based on the idea of marginal risk sets. Considering either clinical or gene expressions while fitting WLW model gives us almost identical results as A-G model. The model (WLW) which considers clinical risk factors only identifies the same covariates as risk factors, i.e., patients' age and patients belong to high pathological stage and the mean ratios are almost the same as we did not include time varying covariates in the model. Though the mean ratios from WLW model and hazard ratios from A-G model give identical estimates, the model which contains gene expressions identifies more genes (CTHRC1, PAGE4, MAPK3, CCL2, ALOX15, SNHG8) as significant risk factors for the recurrence of tumors than the A-G model.

Table 3: Effects of Risk Factors: WLW Model

Risk Factors	Mean Ratio (95% CI)
Age	1.041 (0.948, 1.144)
Clinical Stage (High)	0.337 (0.064, 1.762)
Pathological Stage (High)	4.322^{d} (0.969, 19.263)
Treatment (Not Received)	1.884 (0.234, 15.152)
FAM49B (Altered)	5.961^{c} (1.453, 24.448)
PAGE4 (Altered)	3.602^{c} (1.024, 12.674)
BUB1 (Altered)	15.333^{b} (2.566, 19.597)
MKI67 (Altered)	7.912^{c} (1.024, 26.118)
MAPK3 (Altered)	3.121^{d} (0.869, 11.209)
CTGF (Altered)	3.475^{d} (0.957, 12.613)
CCL2 (Altered)	5.796^{b} (1.608, 20.883)

a : P-value $\leq .001$, b : $.001 <$ P-value $\leq .01$,
c : $.01 <$ P-value $\leq .05$, d : $.05 <$ P-value $\leq .1$

Mean Ratios from the WLW model with the combination of both clinical and gene risk factors are listed in Table 3. We found one noticeable difference between the A-G and the WLW models. The WLW model delivers more genes expressions as statistically significant in the recurrence of tumors than A-G model which illustrates that more genes are responsible in the recurrence of tumors. Furthermore, patients who receive treatment and who belong to high pathological stage have higher risk of recurring tumors.

5.3 Prentice, Williams, and Peterson Total Time (PWP-TT) Model

In fitting the third model, Prentice-Williams-Peterson total time (PWP-TT) (Prentice, Williams and Peterson, 1981), we evaluate the effects of a factor for an event since the entry time in the study. While fitting clinical factors and gene expressions separately, patients' age and pathological stage are found to have significant effects on the recurrence of tumors similar to A-G and WLW models. Five genes- CTHRC1, PAGE4, MAPK3, CCL2 and SNHG8 are found to be significant for developing tumors after initial treatment.

Table 4: Effects of Risk Factors: PWP-TT

Risk Factors	Hazard Ratio (95% CI)
Age	1.041 (0.919, 1.180)
Clinical Stage (High)	0.569 (0.042, 7.574)
Pathological Stage (High)	8.500^{d} (0.770, 13.774)
Treatment (Not Received)	1.809 (0.082, 19.933)
CTHRC1 (Altered)	11.961^{c} (1.069, 33.764)
BUB1 (Altered)	9.648^{b} (3.341, 27.046)
FAM49B (Altered)	6.671^{d} (0.719, 21.812)
CCL2 (Altered)	10.510^{b} (2.002, 41.170)

a : P-value $\leq .001$, b : $.001 <$ P-value $\leq .01$,

c : $.01 <$ P-value $\leq .05$, d : $.05 <$ P-value $\leq .1$

Combined model shows (Table 4) that the patients who belong to high pathological stage have 8.500 times higher risk of having recurring tumor than those at low pathological stage. Patients at low clinical stage are 43.1% less likely to have recurring tumor than those at high clinical stage. Patients' who did not receive any treatment have 1.809 times higher risk of recurring tumor. In addition, patients with altered genes as follows, BUB1, PAGE4, FAM49B, MKI67, CTGF, CCL2 have higher risk of having recurring tumors compared to those having non altered genes.

5.4 Prentice, Williams, and Peterson Gap Time (PWP-GT) Model

Prentice, Williams, and Peterson gap time (PWP-GT) (Prentice, Williams and Peterson, 1981) model evaluates the effect of a factor for a particular event since the time from the previous event. Models with clinical factors and gene expressions separately, reveals that patients' age along with pathological stage (high) are significantly associated with the risk of recurring tumor. Also, five genes (CTHRC1, PAGE4, TTTY15, CCL2, ALOX15) are found to be highly significant among twenty one genes for recurring tumor.

Table 5: Effects of Risk Factors: PWP-GT

Risk Factors	Hazard Ratio (95% CI)
Age	1.076 (0.948, 1.221)
Clinical Stage (High)	0.607 (0.130, 2.819)
Pathological Stage (High)	7.887^{b} (1.789, 34.768)
Treatment (Not Received)	1.678 (0.047, 11.128)
CTHRC1 (Altered)	4.524^{c} (1.277, 11.017)
BUB1 (Altered)	8.832^{d} (0.777, 16.359)
PAGE4 (Altered)	4.296^{d} (0.781, 13.613)
CCL2 (Altered)	7.987^{b} (1.946, 32.776)
FAM49B (Altered)	5.682^{c} (1.031, 21.328)

a : P-value $\leq .001$, b : $.001 <$ P-value $\leq .01$,
c : $.01 <$ P-value $\leq .05$, d : $.05 <$ P-value $\leq .1$

Additionally, combined model (Table 5) illustrates that the recurrence of tumors increases with patients' age, and is higher among patients who belong to high-pathological stage, and did not receive any treatment yet. The risk of recurrence of tumors is also higher for the patients with the following altered genes, CTHRC1, BUB1, PAGE4, CCL2, FAM49B.

5.5 Frailty Model

We consider the most commonly used frailty model, the shared frailty (Amorim and Cai, 2015) with random effects. Patients' age and pathological stage are found to be significant risk factor for recurring tumors. Only two gene expressions (CTHRC1, CCL2) among nineteen top-ranked genes are found to be significantly associated with recurring tumor when we consider frailty models separately for clinical factors and gene expressions.

Table 6: Effects of Clinical Risk Factors: Frailty

Risk Factors	Hazard Ratio (95% CI)
Age	1.035 (0.927, 1.157)
Clinical Stage (High)	0.482 (0.103, 2.255)
Pathological Stage (High)	3.767^{d} (0.967, 14.668)
Treatment (Not Received)	1.529 (0.204, 11.408)
BUB1 (Altered)	9.641^{c} (1.054, 18.124)
FAM49B (Altered)	3.862^{d} (0.946, 15.765)
CCL2 (Altered)	4.230^{d} (0.940, 19.022)

a : P-value $\leq .001$, b : $.001 <$ P-value $\leq .01$,
c : $.01 <$ P-value $\leq .05$, d : $.05 <$ P-value $\leq .1$

Table 6 illustrates that the rate of recurrence increases with age, is higher for patients at high pathological stage and those who did not receive treatment. Patients from high clinical stage influences the risk of recurrence, however it is not significantly associated with a decreased risk of recurrence. Also, three gene factors (BUB1, FAM49B, CCL2) show higher risk of recurring tumors for alteration of these genes.

Table 7: Results of Five Analytical Approaches for Recurrent Events: Clinical Risk Factors

	A-G	WLW	PWP-TT	PWP-GT	Frailty
Effects	HR (95% CI)	MR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Age	1.041	1.041	1.041	1.076	1.035
	(0.941, 1.152)	(0.948, 1.144)	(0.919, 1.180)	(0.948, 1.221)	(0.927, 1.157)
Clinical Stage	0.337	0.337	0.569	0.607	0.482
(High)	(0.051, 2.242)	(0.064, 1.762)	(0.042, 7.574)	(0.130, 2.819)	(0.103, 2.255)
Pathological Stage	$\mathbf{4.322}^{d}$	$\mathbf{4.322}^{d}$	$\mathbf{8.500}^{d}$	$\mathbf{7.887}^{b}$	$\mathbf{3.767}^{d}$
(High)	(0.839, 22.259)	(0.969, 19.263)	(0.770, 13.774)	(1.789, 34.768)	(0.967, 14.668)
Treatment	1.884	1.884	1.809	1.678	1.529
(Not Received)	(0.186, 19.062)	(0.234, 15.152)	(0.082, 19.933)	(0.047, 11.128)	(0.204, 11.408)

a : P-value $\leq .001$, b : $.001 <$ P-value $\leq .01$, c : $.01 <$ P-value $\leq .05$, d : $.05 <$ P-value $\leq .1$

Finally Tables 7 and 8 summarize the hazard ratios (HR) or mean ratios (MR) and corresponding 95% confidence intervals for the combined risk factors for each of the five models considered in this study. Table 7 presents the clinical risk factors. The risk of

recurring tumors increases with patients' age but is not statistically significant. Treatment is another risk factor which is also not statistically significant. Nevertheless, the group of patients who did not receive treatment has higher hazard of recurring tumors compared to the group who received treatment. Patients in high pathological stage are significantly associated with the recurrence, i.e., patients who are already in the high pathological stage are prone to the recurrence of tumor and the rate is higher compared to low pathological stage patients for all five models.

Table 8: Results of Five Analytical Approaches for Recurrent Events: Gene Expressions

	A-G	WLW	PWP-TT	PWP-GT	Frailty
Effects	HR (95% CI)	MR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
BUB1	**15.333**^{c}	**15.333**^{b}	**9.648**^{b}	**8.832**^{d}	**9.641**^{c}
	(1.784, 31.736)	(2.566, 19.597)	(3.341, 27.046)	(0.777, 16.359)	(1.054, 18.124)
PAGE4	**3.602**^{d}	**3.602**^{c}	3.351	**4.296**^{d}	3.107
	(0.885, 14.662)	(1.024, 12.674)	(0.590, 19.021)	(0.781, 13.613)	(0.740, 13.046)
FAM49B	**5.961**^{d}	**5.961**^{c}	**6.671**^{d}	**5.682**^{c}	**3.862**^{d}
	(1.146, 30.982)	(1.453, 24.448)	(0.719, 21.812)	(1.031, 21.328)	(0.946, 15.765)
MKI67	**7.912**^{d}	**7.912**^{c}	9.335	7.708	4.972
	(0.830, 17.399)	(1.024, 16.118)	(0.607, 15.443)	(0.406, 17.435)	(0.350, 17.580)
CTGF	**3.475**^{d}	**3.475**^{d}	3.650	-	2.062
	(0.790, 15.279)	(0.957, 12.613)	(0.552, 14.120)	-	(0.292, 14.556)
CCL2	**5.796**^{c}	**5.796**^{b}	**10.510**^{b}	**7.987**^{b}	**4.230**^{d}
	(1.359, 24.704)	(1.608, 20.883)	(2.002, 41.170)	(1.946, 32.776)	(0.940, 19.022)
MAPK3	3.121	**3.121**^{d}	3.932	-	2.493
	(0.671, 11.499)	(0.869, 11.209)	(0.570, 9.120)	-	(0.662, 9.393)
CTHRC1	4.003	4.003	**9.961**^{c}	**4.524**^{c}	3.043
	(0.559, 10.078)	(0.716, 9.096)	(1.069, 13.076)	(1.277, 11.017)	(0.650, 14.236)

a : P-value $\leq .001$, b : $.001 <$ P-value $\leq .01$, c : $.01 <$ P-value $\leq .05$, d : $.05 <$ P-value $\leq .1$

Table 8 compares the effects of gene expressions controlling for the clinical factors using five models. In the absence of time dependent covariates (patients' age was recorded during the first diagnosis of prostate cancer), A-G and WLW model produce identical point estimates for hazard and mean ratios. However, the confidence intervals are different for these two models due to their distinct estimation procedure. Estimates from both PWP models are based on risk sets which only include those with the same number of prior events. However, for each statistically significant gene expression, frailty model produces comparatively lower hazard rate than other models. One plausible reason is that frailty models assume that

patients' susceptibility to risk of cancer recurrence differs and the induced correlation among the recurrent events are addressed explicitly; whereas other four models do not address the heterogeneous susceptibility to the risk of recurrent events. Three genes are found to be statistically significant for recurring tumors in all five models (BUB1, FAM49B, CCL2). It is to be noted that all four models except PWP-GT found the hazards for recurrent tumors associated with altered gene BUB1 as statistically significant. All four models except frailty model found the hazards for recurrent tumors associated with altered gene CCL2 as statistically significant. Whereas three models, A-G, WLW and PWP-GT found the hazards for recurrent tumors associated with altered gene FAM49B as statistically significant. Thus these genes may be associated with higher risk of recurring tumors among prostate cancer patients.

6 Conclusion

The primary objective of this study is to identify the risk factors for the recurrent time to tumor events on prostate cancer patients and compare their effects obtained from different statistical approaches for analyzing recurrent time to event data. In this study, we consider five commonly used extended Cox models for recurrent events, namely, A-G, WLW, PWP-TT, PWP-GT and Frailty model. These five models differ in assumptions as well as the mechanism of modeling recurrent events. A-G and PWP model assume that any particular event depends only on prior event. WLW model assumes no specific dependence structures among the recurrent event times within a subject, and distinguishes the mean or rates of the counting process. Frailty model assumes dependency among the recurrent event times within a subject through shared random effects.

Analysis of recurrence of tumors among prostate cancer patients using five statistical models for recurrent events results in somewhat similar clinical and genetics factors. The clinical factors which are found to be influential on the survival of the prostate cancer patients in the literature are also found to be important for recurrence of tumors after the diagnosis. Elderly men are prone to have recurring tumors and the risk of older age is found to be high in all five recurrent event models. The hazards for having recurring tumors among patients in high pathological stage is remarkably greater compared to those in lower stage, although the hazard ratio is not statistically significant in four of the five models. Only PWP-GT model produce a relatively higher and statistically significant hazard ratio for high pathological stage patients. Nonetheless these results illustrate that patients may have recurring tumors even after surgery. Therapeutic treatment (any of the molecular, neoadjuvant, chemotherapy or radiation therapy) is an important factor, and the results from all the five models illustrate that the patients who had not received any treatment are about twice as likely to have recurrent tumors compared to those received treatment.

When considered separately a number of gene's expression status (altered versus not-altered) are found to be associated with the recurrence of tumors from all five models for recurrent events. However, after controlling for clinical factors less number of genes become

influential in recurring of tumors among the prostate cancer patients. Furthermore, these influential genes have higher risk of tumor recurrence if they belong to the altered group whose standardized expression counts are greater than the threshold value of 2. Three genes are found to be statistically significant for recurring tumors in all five models (BUB1, FAM49B, CCL2). All four models except PWP-GT found the hazards for recurrent tumors associated with altered gene BUB1 as statistically significant. All four models except frailty model found the hazards for recurrent tumors associated with altered gene CCL2 as statistically significant. Whereas three models, A-G, WLW and PWP-GT found the hazards for recurrent tumors associated with altered gene FAM49B as statistically significant. Thus it is reasonable to conclude that these genes may be associated with higher risk of recurring tumors among prostate cancer patients.

Missingness is a major drawback in biomedical studies. In particular, often times missing data fail to observe and include all important information about the patients. We completely ignored the missingness of the covariates and excluded one covariate (e.g., patient's race) from the model which was found significant on previous studies due to excessive missing values. Taking into account the covariates for which information was incomplete might give us more accurate results.

In this research, we considered five models (A-G, WLW, PWP-TT, PWP-GT and Frailty) to analyze recurrent tumors among prostate cancer patients as we limited our study considering time independent variables as well as few recurrence per subject. Inclusion of time dependent covariates, incomplete covariate information and consideration of other approaches for recurrent events are left as future research. In addition, similar approaches can be implemented to other prominent cancer data from TCGA.

Acknowledgment

The authors would like to thank three anonymous referees for their critical readings and most helpful comments which improve the paper significantly.

References

Amorim, L. D., Cai, J. (2015), "Modelling recurrent events: a tutorial for analysis in epidemiology", *International Journal of Epidemiology*, 44(1), 324–333.

Andersen, P. K., Gill, R. D (1982), "Testing goodness of fit of Coxs regression and life model", *Biometrics*, 38(0), 67–77.

Barbieri, C. E., Baca, S. C. (2012), "Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer", *Nature: Genetics*, 44, 685-689.

Burns, R. M., Sharp, L., Sullivan, F. J., et al. (2014), "Factors Driving Inequality in Prostate Cancer Survival: A Population Based Study", *PMC*, 9(9), e106456.

Chan, Y. M.(2013), "Statistical Analysis and Modeling of Prostate Cancer", *Thesis Dissertation: University of South Florida.*

Clayton, D. and Cuzick, J. (1985). "Multivariate generalizations of the proportional hazards model", *Journal of the Royal Statistical Society, Series A*, 148(2), 82–117.

Jhun, M. A., Geybels, M. S., Wright, J. L., et al. (2017), "Gene expression signature of Gleason score is associated with prostate cancer outcomes in a radical prostatectomy cohort", *Oncotarget*, 8(26), 43035–43047.

Kelly, P. J., Lim, L. L-Y. (2000), "Survival analysis for recurrent event data: an application to childhood infectious diseases", *Statistics in Medicine*, 19, 13–33.

Kenfield, S. A., Stampfer, M. J., Chan, J. M., Giovannucci, E. (2011), "Smoking and prostate cancer survival and recurrence", *JAMA*, 305(24), 2548-2555.

Liu, L., Wolfe, R. A., Huang, X. (2004), "Shared frailty models for recurrent events and a terminal event", *Biometrics*, 60, 747-756.

Mortensen, M. M., Hoyer, S., Lynnerup, A. S., et al. (2015), "Expression profiling of prostate cancer tissue delineates genes associated with recurrence after prostatectomy", *European Urology Focus*, 5, 16018.

Nielsen, G. G., Gill, R. D., Andersen, P. K., Sorensen, T. I. A. (1992), "A counting process approach to maximum likelihood estimation in frailty models", *Scandinavian Journal of Statistics*, 19(1), 25–43.

Oakes, D. (1992), *Survival analysis: state of the art*, Springer, Netherlands.

Paul, L. A (2010), "Statistical design and analysis of next generation sequencing data", *Thesis Dissertation: Purdue University.*

Prentice, R. L., Williams, B. J., Peterson, A. V. (1981), "On the regression analysis of multivariate failure time data", *Biometrika*, 68(2), 373–379.

Robinson, D., et al. (2015), "Integrative clinical genomics of advanced prostate cancer", *Science Direct: Cell*, 161(5), 1215–1228.

Ross, J. S., Sheehan, C. E., Fisher, H. A. G., et al. (2003), "Correlation of primary tumor prostate-specific membrane antigen expression with disease recurrence in prostate cancer", *Clinical Cancer Research*, 9(17), 6357–6362.

Shancheng, R., Zhiyu, P., et al. (2012), "RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings", *Nature: Cell Research*, 22, 806-821.

Ullah, S., Gabbett, T. J., Finch, C. F. (2012), "Statistical modelling for recurrent events: an application to sports injuries", *British Journal of Sports Medicine*, published online.

Wei, L. J., Lin, D. Y., Weissfeld, L. (1989), "Regression analysis of multivariate incomplete failure time data by modeling marginal distributions", *Journal of the American Statistical Association*, 84(408), 1065–1073.

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

Wilding, G., Remington, P. (2005), "Period analysis of prostate cancer survival", *Journal of Clinical Oncology*, 23(3), 407–409.

William, J. L., Greer, P. A., Squire, J. A. (2014), "Recurrent copy number alterations in prostate cancer: an in silico meta-analysis of publicly available genomic data", *Science Direct: Cancer Genetics*, 207(1012), 474–488.

Zare, A., Hossaini, M., Mahmoodi, M., Mohammad, K., et al. (2015), "A comparison between accelerated failure-time and cox proportional hazard models in analyzing the survival of gastric cancer patients", *Iranian Journal of Public Health*, 44(8), 1095–1102.

Received: July 12, 2017

Accepted: January 7, 2018