Journal of Data Science 15(2017), 423-442

# COMPARISON OF COX REGRESSION AND PARAMETRIC MODELS FOR SURVIVAL ANALYSIS OF GENETIC VARIANTS IN HNF1B GENE RELATED TO AGE AT ONSET OF CANCER

Kesheng Wang[1],*, Xuefeng Liu[2], Yue Pan[3], Daniel Owusu[1], Chun Xu[4]

[1]*Department of Biostatistics and Epidemiology, East Tennessee State University, Johnson City, TN 37614, USA*

[2]*Department of Systems Leadership and Effectiveness Science, University of Michigan, Ann Arbor, MI 48109-5482, USA*

[3]*Department of Public Health Sciences, Miller School of Medicine, University of Miami, Miami, FL 33136, USA*

[4]*Department of Health and Biomedical Sciences, University of Texas Rio Grande Valley, Brownsville, TX 78520, USA*

*Abstract:* Semi-parametric Cox regression and parametric methods have been used to analyze survival data of cancer; however, no study has focused on the comparison of survival models in genetic association analysis of age at onset (AAO) of cancer. The Hepatocyte nuclear factor-1-beta (HNF1B) gene has been associated with risk of endometrial and prostate cancers; however, no study has focused on the effect of HNF1B gene on the AAO of cancer. This study examined 23 single nucleotide polymorphisms (SNPs) within the HNF1B gene in the Marshfield sample with 716 cancer cases and 2,848 non-cancer controls. Cox proportional hazards models in PROC PHREG and parametric survival models (including exponential, Weibull, log-normal, log-logistic, and gamma models) in PROC LIFEREG in SAS 9.4 were used to detect the genetic association of HNF1B gene with the AAO. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to compare the Cox models and parametric survival models. Both AIC and BIC values showed that the Weibull distribution is the best model for all the 23 SNPs and the Gamma distribution is the second best. The top two SNPs are rs4239217 and rs7501939 with time ratio (TR) =1.08 ($p<0.0001$ for the AA and AG genotypes, respectively) and 1.07 ($p=0.0004$ and 0.0002 for CC and CT genotypes, respectively) based on the Weibull model, respectively. This study shows that the parametric Weibull distribution is the best model for the genetic association of AAO of cancer and provides the first evidence of several genetic variants within the HNF1B gene associated with AAO of cancer.

*Key words*: Cancer, age at onset, HNF1B, SNP, survival analysis, Cox regression, parametric models

## 1. Introduction

Survival analysis methods, including non-parametric Kaplan-Meier method (including log-rank test and Wilcoxon test), semi-parametric Cox proportional hazards model as well as parametric methods (such as exponential, Weibull, gamma, log-normal, and log-logistic models) have been used in cancer survival studies. However, previous studies have shown inconsistent results for survival analysis methods in cancer survival studies. For example, the Cox model is similar to the exponential model (Pourhoseingholi et al., 2007); while the Weibull and exponential models were similarly the best models in the survival analysis of stomach cancer (Moghimi-Dehkordi et al., 2008). In other studies, the Weibull model was shown to be better than the

exponential model in survival analysis of patients with gastric cancer in Iran (Baghestani et al., 2009), and was the best model for patients with gastric cancer (Zhu et al., 2011) and colorectal cancer in Iran (Baghestani et al., 2015). The log-normal survival model may have a good fit for the gallbladder cancer dataset (Wang et al., 2010); while the log-logistic model with gamma frailty is the better-fit model in the survival analysis of gastrointestinal cancer in northern Iran (Ghadimi et al., 2011). Another study has shown parametric models such as Weibull model, log-normal and gamma models may perform better than Cox models in identifying prognostic factors for oral cancer (Köhler and Kowalski, 2012).

Family and twin studies have indicated that genetic and environmental factors and their interactions contribute to the development of human cancers, with a heritability of 22% for ovarian cancer, 35% for colorectal cancer (Lichtenstein et al., 2000), 25-30% for breast cancer (Lichtenstein et al., 2000; Czene et al., 2002; Locatelli et al., 2004), and 42-58% for prostate cancer (Lichtenstein et al., 2000; Hjelmborg et al., 2014). The Hepatocyte nuclear factor-1-beta (HNF1B) gene (also known as FJHN; HNF2; LFB3; TCF2; HPC11; LF-B3; MODY5; TCF-2; VHNF1; HNF-1B; HNF1beta) is located at 17q12 (Abbott et al., 1990; Bach et al., 1991; Gudmundsson et al., 2007) and a member of the homeodomain-containing superfamily of transcription factors (Bach et al., 1991). HNF1B may play a role in renal cell carcinoma (Rebouissou et al., 2005), ovarian cancer, as well as gastric, pancreatic, and colorectal cell lines (Terasawa et al., 2006). Recently, several candidate gene and genome-wide association studies (GWAS) reported that several single nucleotide polymorphisms (SNPs such as rs7501939 and rs4430796) in the HNF1B gene were associated with the risks of endometrial and prostate cancers (Gudmundsson et al., 2007; Spurdle et al., 2011; Wang et al., 2014). Age at onset (AAO) of complex diseases such as cancers have also genetic components (e.g., Claus ., 1990; Shugart et al., 2000; Pankratz et al., 2005; Reeves et al., 2009; Chen et al., 2013); however, no study has examined the effect of HNF1B gene on AAO of cancer. The semi-parametric Cox proportional hazards model has been used to examine the associations of genetic variants with AAO of cancer. For example, previous studies tested the association between p53 and DNMT3b polymorphisms and AAO of colorectal cancer using the Cox proportional hazards regression model (Jones et al., 2004; Krüger et al., 2005; Sotamaa et al., 2005; Jones et al., 2006). Furthermore, Kulminski et al. (2011) reported the association between apolipoprotein E (APOE) e2/3/4 polymorphism and AAO of cancer (all sites but skin) using data on 3924 participants of the Framingham Heart Study Offspring cohort in Cox regression model; while Chen et al. (2013) examined the association of 1456 SNPs in 128 cell cycle-related genes and 31 DNA repair-related genes in 485 non-Hispanic Whites to determine whether there were SNPs associated with AAO of colorectal cancer. Recently, Wang et al. (2015) examined the associations of 220 SNPs within the PTPRN2 gene with the AAO of cancer using the Cox regression. However, no study has compared the survival models in genetic association analysis of age at onset (AAO) of cancer.

This study was to identify the best model by comparing the Cox proportional hazards models and parametric survival models in genetic association analysis of the HNF1B gene with the AAO of cancer in a Caucasian sample.

## 2. Subjects and Methods

### 2.1. The Marshfield sample

The Marshfield sample was selected from the publicly available data in A Genome-Wide Association Study on Cataract and HDL in the Personalized Medicine Research Project Cohort - Study Accession: phs000170.v1.p1 (dbGaP). Details about the participants were described elsewhere (McCarty et al., 2005, 2008). Cases were defined as any diagnosed cancer excluding minor skin cancer and AAO of cancer was defined by the date of the earliest cancer diagnosis in the registry. Covariates

included in this study were age, gender, alcohol use in the past month (yes or no), obesity status, and smoking status (never smoking, current smoking and past smoking). Obesity was defind as a body mass index (BMI) $\geq$ 30 kg/m$^2$. Genotyping data using the ILLUMINA Human660W-Quad_v1_A were available for 3894 individuals. The genotypes of 23 SNPs within the HNF1B gene were available in this data.

## 2.2. Descriptive Statistics and Quality Control

Categorical variables were presented as frequencies and percentages, while continuous variables were reported as the means ± standard deviation (SD). HelixTree Software (http://www.goldenhelix.com/SNP_Variation/HelixTree/index.html) was used to assess control genotype data for conformity with Hardy-Weinberg equilibrium (HWE). Genotype call rates and minor allele frequency (MAF) were also calculated. To account for population stratification, the principal-component analysis approach (Price et al., 2006) in HelixTree software was used to identify outlier individuals (Wang et al., 2012). Based on the principal components analysis of the first 5 principal components using HelixTree and genome-wide genotype data, we removed outlier individuals. Consequently, 3564 Caucasian individuals were included in the analysis (716 cancer cases and 2848 controls).

## 2.3. Cox Proportional Hazards Model

The proportional hazards model or Cox regression model (Cox, 1972) is widely used in the analysis of time-to-event data to explain the effect of explanatory variables on hazard rates (Cantor, 2007; George et al., 2014).

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p x_p) \quad (1)$$

where $h(t|\boldsymbol{x})$ is the hazard function at time t for a subject with a set of predictors x1,…,xp, h0(t) is the baseline hazard function, and β1,…,βp are the model parameters describing the effect of the predictors on the overall hazard. Then the hazard ratio (HR) is defined as the ratio of predicated hazard rates under two different values of a predictor variable (George et al., 2014). The PHREG procedure in SAS was used to fit the Cox regression model by considering censoring and maximizing the partial likelihood function.

## 2.4. Parametric Survival Models

Some commonly assumed parametric distributions in survival models include exponential, Weibull, gamma, log-normal, and log-logistic (Klein and Moeschberger, 2003).

$$\ln(T) = \beta_1 x_1 + \ldots + \beta_p x_p + ln(\varepsilon) \quad (2)$$

where $T$ is the time to event; x1,…,xp, and β1,…,βp are predictor variables and their corresponding coefficients, respectively; ε is the error term assumed to have a particular parametric distribution; and ln(ε) is the natural log of the error term (George et al., 2014). The exponentials of the β coefficients may be interpreted as the time ratio (TR) (Hernán et al., 2005; George et al., 2014; Kasza et al., 2014). If TR >1, the event is less likely to occur as it means it will take longer for the event to happen; whereas if TR <1, the event is more likely to happen. The LIFEREG procedure in SAS fits parametric survival models, where the link function can be taken from a class of distributions that include exponential, Weibull, log-normal, log-logistic, and gamma distributions.

### 2.5. Evaluation Criteria for Goodness of Fit

The Akaike information criterion (AIC) was used as a measure of goodness of model fit that balances model fit against model simplicity (Akaike, 1979, 1981); while the Bayesian information criterion (BIC) was used as a similar measure (Simonoff, 2003).

$$\text{AIC} = -2\ln\{p(x|\hat{\theta})\} + 2k \quad (3)$$

and

$$\text{BIC} = -2\ln\{p(x|\hat{\theta})\} + k\ln n \quad (4)$$

where x is the random variable, $\hat{\theta}$ is the maximum likelihood estimate, k is the number of parameters, and n is the sample size. Note that model with smaller AIC and BIC values fits the data better.

### 2.6. Survival Analysis of Age at Onset of Cancer

The assessment of the association between genotypes of each SNP and AAO was initially performed using the log-rank test and Wilcoxon test in Kaplan–Meier (KM) survival analysis using LIFETEST procedure. The KM survival curves were used to plot the survival function. The PHREG procedure in SAS was used to fit the Cox model while the LIFEREG procedure was used to fit parametric survival models including the exponential, Weibull, log-normal, log-logistic, and gamma distributions. Multivariate Cox regression analysis and parametric survival analyses were conducted to detect associations of each SNP with AAO adjusting for gender, alcohol use in the past month, smoking status and obesity status. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to compare the Cox regression and parametric survival models. Descriptive statistics, the KM survival analysis, Cox regression, and parametric model analyses were conducted with SAS v.9.4 (SAS Institute, Cary, NC, USA). SAS codes are listed in Appendix.

## 3. Results

### 3.1. Genotype Quality Control and Descriptive Statistics

All the 23 SNPs were in HWE in the controls ($p>0.001$) with MAF >5%. The demographic characteristics of the subjects in the study are presented in Table 1. There were slightly more females than males in both cases and controls. Age ranged from 46 to 90 years and AAO of cancer ranged from 23 to 90 years.

### 3.2. Comparison of Cox Regression and Parametric Models using PROC PHREG and PROC LIFEREG

The estimates of AIC and BIC consistently showed that the Weibull distribution was the best model for all the 23 SNPs and the Gamma distribution was the second best. Table 2 shows AIC and BIC for the different models for 5 SNPs associated with AAO in the Weibull model ($p<0.05$). The top two SNPs were rs4239217 and rs7501939, in that order. For rs4239217, the Weibull distribution was the best model (AIC and BIC are 5582.38 and 5623.38, respectively) and the Gamma distribution was the second (AIC and BIC are 5583.94 and 5629.6, respectively). For the SNP rs7501939, the AIC and BIC values showed consistent results that the Weibull distribution was the best model (AIC and BIC are 5582.38 and 5623.38, respectively) and the Gamma model was the second best (AIC and BIC are 5583.94 and 5629.6, respectively).

### 3.3. Survival Analysis of Age at Onset using the Weibull Model

The results of the parametric survival analysis using the Weibull model is presented in Table 3. Five SNPs (rs3110649, rs757210, rs430796, rs4239217 and rs7501939) showed associations with AAO of cancers. The top SNP was rs4239217 with TR=1.08 and the second signal was rs7501939 with TR=1.07. The Kaplan–Meier survival curves for different genotypes of SNPs rs4239217 and rs7501939 are shown in Figures 1 and 2, respectively. For rs4239217, the mean AAO was approximately 5.5 and 4.6 years later for individuals with AA genotype and AG genotype compared to those with GG genotype ($p=0.0002$ based on the log-rank test and $p<0.0001$ based on the Wilcoxon test). For rs7501939, the mean AAO was approximately 5.2 and 4.3 years later for individuals with CC genotype and CT genotype compared to those with TT genotype ($p=0.0058$ based on the log-rank test and $p=0.0002$ based on Wilcoxon test).

## 4. Discussion

In the present study, we explored the associations of 23 HNF1B SNPs with the AAO of cancer. Using the AIC and BIC, the Weibull distribution was found to be the best model for genetic association of polymorphisms within the HNF1B gene with the AAO of cancer. To our knowledge, this is the first study to compare the Cox regression and parametric survival models in genetic association analysis of AAO of cancer. It is also the first candidate gene study to provide evidence of several genetic variants (rs3110649, rs757210, rs430796, rs4239217 and rs7501939) within the HNF1B gene which may be involved in the AAO of cancer.

Semi-parametric Cox regression and parametric survival methods (such as exponential, Weibull, gamma, log-normal, and log-logistic models) have been used in cancer survival studies. However, previous studies have shown inconsistent results for survival analyses in cancers. For example, one study showed that the Cox regression is similar to the exponential model (Pourhoseingholi et al., 2007); while another study showed parametric models such as Weibull model, lognormal and gamma models may perform better than Cox model in oral cancer (Köhler and Kowalski, 2012). Furthermore, some studies favored the Weibull model in stomach cancer (Moghimi-Dehkordi et al., 2008), gastric cancer (Baghestani et al., 2009; Zhu et al., 2011) and colorectal cancer (Baghestani et al., 2015). In addition, one study found that the log-normal survival model may have a good fit for the gallbladder cancer (Wang et al., 2010); while the log-logistic model with gamma frailty is the best model in gastrointestinal cancer in northern Iran (Ghadimi et al., 2011). Several studies have used the semi-parametric Cox model to examine the association of genetic variants with AAO of colorectal cancer (Jones et al., 2004; Krüger et al., 2005; Sotamaa et al., 2005; Jones et al., 2006; Chen et al., 2013) and of cancer (all sites but skin) (Kulminski et al., 2011; Wang et al., 2015). However, no study was found to have examined associations between genetic variants and AAO of cancer using parametric models (including exponential, Weibull, log-normal, log-logistic and gamma models). Therefore, the current study is the first attempt to compare the Cox regression with parametric survival models in genetic association analysis of AAO of cancer. Furthermore, in consistent with some previous studies (such as Moghimi-Dehkordi et al., 2008; Baghestani et al., 2009; Zhu et al., 2011; Baghestani et al., 2015), our results showed that the Weibull distribution is the best model for genetic associations of all 23 SNPs within the HNF1B gene with AAO of cancer. Moreover, we found that the Gamma distribution is the second best model. The differences of these comparisons may be due to different cancer types, sample size, and sample origins. On the other hand, we performed genetic association study of the AAO of cancer using parametric survival models.

It was shown that HNF1B regulated the expression of polycystic kidney and hepatic disease-1 (PKHD1) and therefore HNF1B may function as a tumor suppressor gene in chromophobe renal cell carcinogenesis (Rebouissou et al., 2005); while HNF1B may be involved in the development of ovarian cancers, gastric, pancreatic, and colorectal cell lines (Terasawa et al., 2006). Recently, several candidate genes and GWAS studies observed that rs4430796 within HNF1B gene was associated with prostate cancers (Gudmundsson et al., 2007; Eeles et al., 2008; Levin et al., 2008; Thomas et al., 2008; Waters et al., 2009; Berndt et al., 2011; Spurdle et al., 2011, Wang et al., 2014) and endometrial cancer (Spurdle et al., 2011; Wang et al., 2014). The

signal of rs7501939 was reported to be associated with prostate cancers (Eeles et al., 2008; Levin et al., 2008; Wang et al., 2014) and endometrial cancer (Setiawan et al., 2012; Wang et al., 2014). Several studies also found that rs4430796 was associated with lung cancer in Chinese population (Sun et al., 2011), and prostate cancer in Korean men (Kim et al., 2008), Chinese men (Zhang et al., 2012), and African American men (Chornokur et al., 2013). However, no study has focused on the effect of HNF1B gene on AAO of cancer. In the present study, we provided the first evidence that two previously cancer risk associated SNPs (rs430796 and rs7501939) within HNF1B gene were associated with AAO of cancer. We added that 3 more SNPs (rs3110649, rs757210, and rs4239217) were associated with AAO of cancer.

Studies suggest that type 2 diabetes (T2D) might share the same genetic link to prostate cancer. HNF1A S319 was indicated to be associated with earlier AAO of T2D in women (Hegele et al., 2000). Two SNPs (rs7501939 and rs4430796) of HNF1B were reported to be associated with T2D in Chinese as well as in Caucasians (Gudmundsson et al., 2007). Recently, the results of some GWAS studies provided support for a shared genetic contribution to the risk of T2D and prostate cancer. For example, in the study by Gudmundsson et al. (2007), the A allele of rs4430796 and C allele of rs7501939 variants in HNF1B/TCF2 showed positive associations with prostate cancer (OR>1.0) but was protective against T2D (OR<1.0). A later study confirmed the association of rs4430796 with T2D and prostate cancer and suggested that T2D had a protective effect on prostate cancer risk (Piece et al., 2010). A meta-analysis examined the two variants (rs4430796 and rs7501939) and found they had pleiotropic effects on T2D and prostate cancer (Elliott et al., 2010). The present study added that the HNF1B gene may play a role in the development of cancer.

There are some limitations in this study. First, the definition of cancer status in the Marshfield sample was broad (including any diagnosed cancer omitting minor skin cancer) which may result in genetic and phenotypic heterogeneity into the genetic association analysis. It would be more informative to investigate the association of HNF1B with specific types of cancer. Second, our current findings might be subject to type I error and findings need to be replicated in additional samples. In addition, in the present study, we just used the original Cox regression model and parametric survival models. Interestingly, investigators have extended original Cox regression and parametric models. For example, Li et al. (2016) recently developed the proportional generalized odds (PGO) model, which covers the proportional odds (PO) model (Bennett, 1983; Pettitt. 1984) and the generalized proportional odds (GPO) model (Dabrowska and Doksum, 1988). On the other hand, Musrafa et al. (2016) proposed a new four parameters Weibull model called the Weibull Generalized Flexible Weibull extension (WGFWE) distribution and Alkarni (2016) introduced a new family of models for lifetime data called generalized extended Weibull power series family of distributions by compounding generalized extended Weibull distributions and power series distributions; while Pu et al. (2016) proposed a new class of five parameters gamma-exponentiated or generalized modified Weibull (GEMW) distribution. In the present study, we just tested two parameters Weibull model. In the future, it will be prospective to test and apply these extended survival models in the genetic association of AAO of complex diseases.

There are also several strengths in this study. First, our sample size was relatively large for this type of study. Second, we compared the semi-parametric Cox regression and parametric models in the genetic association of AAO of cancer. Third, we examined 23 SNPs within the HNF1B gene and especially identified 2 cancer and T2D associated SNPs (rs4430796 and rs7501939) influencing the AAO of cancer.

In conclusion, the results demonstrate that the parametric Weibull model performed better than Cox regression and other parametric models (including exponential, log-normal, log-logistic and gamma models) for the genetic association of AAO of cancer. Furthermore, this study provides evidence of several genetic variants within the HNF1B gene influencing AAO of cancer. These findings may serve as a resource for replication in other populations. Future functional study of this gene may help to better characterize the genetic architecture of the AAO of cancer.

## Acknowledgement

Funding support for the Personalized Medicine Research Project (PMRP) was provided through a cooperative agreement (U01HG004608) with the National Human Genome Research Institute (NHGRI), with additional funding from the National Institute for General Medical Sciences (NIGMS). The samples used for PMRP analyses were obtained with funding from Marshfield Clinic, Health Resources Service Administration Office of Rural Health Policy grant number D1A RH00025, and Wisconsin Department of Commerce Technology Development Fund contract number TDF FYO10718. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Assistance with phenotype harmonization and genotype cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000170.v1.p1. This study was approved by the Internal Review Board (IRB), East Tennessee State University.

## Role of the funding sources

No founding source is given for the present paper.

## Conflict of interest

All authors have reported no financial interests or potential conflicts of interest.

## References

[1] Abbott, C., Piaggio, G., Ammendola, R., Solomon, E., Povey, S., Gounari, F., De Simone, V., Cortese, R. (1990). Mapping of the gene TCF2 for the transcription factor LFB3 to human chromosome 17 by polymerase chain reaction. Genomics 8, 165-167.

[2] Akaike, H. (1979). A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. Biometrika 66, 237–242.

[3] Akaike, H. (1981). Likelihood of a Model and Information Criteria. Journal of Econometrics 16, 3–14.

[4] Alkarni, S.H. (2016). Generalized extended Weibull power series family of distributions. J Data Sci, 14, 415-440.

[5] Bach, I., Mattei, M.-G., Cereghini, S., Yaniv, M. (1991). Two members of an HNF1 homeoprotein family are expressed in human liver. Nucleic Acids Res 19, 3553-3559.

[6] Baghestani, A.R., Gohari, M.R., Orooji, A., Pourhoseingholi, M.A., Zali, M.R. (2015). Evaluation of parametric models by the prediction error in colorectal cancer survival analysis. Gastroenterol Hepatol Bed Bench 8(3), 183-7.

[7] Baghestani, A.R., Hajizadeh, E., Fatemi, S.R. (2009). Bayesian analysis for survival of patients with gastric cancer in Iran. Asian Pac J Cancer Prev 10(5), 823-6.

[8] Bennett, S. (1983). Analysis of survival data by the proportional odds model. Statistics in Medicine 2, 273-277.

[9] Berndt, S.I., Sampson, J., Yeager, M., Jacobs, K.B., Wang, Z., Hutchinson, A., Chung, C., Orr, N., Wacholder, S., Chatterjee, N., Yu, K., Kraft, P., Feigelson, H.S., Thun, M.J., Diver, W.R., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F.R., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G.L., Crawford, E.D., Haiman, C., Henderson, B., Kolonel, L., Le Marchand, L., Siddiq, A., Riboli, E., Travis, R.C., Kaaks, R., Isaacs, W., Isaacs, S., Wiley, K.E., Gronberg, H., Wiklund, F., Stattin, P., Xu, J., Zheng, S.L., Sun, J., Vatten, L.J., Hveem, K., Njølstad, I., Gerhard, D.S., Tucker, M., Hayes, R.B., Hoover, R.N., Fraumeni, J.F. Jr,, Hunter, D.J., Thomas, G., Chanock, S.J. (2011). Large-scale fine mapping of the HNF1B locus and prostate cancer risk. Hum Molec Genet 120, 3322-3329.

[10] Cantor, A.B. (2007). SAS Survival Analysis Techniques for Medical Research. Third Edition. Cary, NC: SAS institute INC.

[11] Chen, J., Pande, M., Huang, Y.J., Wei, C., Amos, C.I., Talseth-Palmer, B.A., Meldrum, C.J., Chen, W.V., Gorlov, I.P., Lynch, P.M., Scott, R.J., Frazier, M.L. (2013). Cell cycle-related genes as modifiers of age of onset of colorectal cancer in Lynch syndrome: a large-scale study in non-Hispanic white patients. Carcinogenesis 34(2),299-306.

[12] Chornokur, G., Amankwah, E.K., Davis, S.N., Phelan, C.M., Park, J.Y., Pow-Sang, J., Kumar, N.B. (2013). Variation in HNF1B and Obesity May Influence Prostate Cancer Risk in African American Men: A Pilot Study. Prostate Cancer 2013, 384594.

[13] Claus, E.B., Risch, N.J., Thompson, W.D. (1990). Using age of onset to distinguish between subforms of breast cancer. Ann Hum Genet 54(Pt 2), 169-77.

[14] Cox, D.R. (1972). Regression models and life-tables. J R Stat Soc Ser B (Methodol) 34(2),187–220.

[15] Czene ,K., Lichtenstein, P., Hemminki, K. (2002). Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. Int J Cancer 99(2), 260-6.

[16] Dabrowska, M. D. and Doksum, K. A. (1988). Estimation and testing in a two-sample generalized odds-rate model. JASA 83, 744-749.

[17] Eeles, R. A., Kote-Jarai, Z., Giles, G. G., Al Olama, A. A., Guy, M., Jugurnauth, S. K., Mulholland, S., Leongamornlert, D.A., Edwards, S.M., Morrison, J., Field, H.I., Southey, M.C., Severi, G., Donovan, J.L., Hamdy, F.C., Dearnaley, D.P., Muir, K.R., Smith, C., Bagnato, M., Ardern-Jones, A.T., Hall, A.L., O'Brien, L.T., Gehr-Swain, B.N., Wilkinson, R.A., Cox, A., Lewis, S., Brown, P.M., Jhavar, S.G., Tymrakiewicz, M., Lophatananon, A., Bryant, S.L.; UK Genetic Prostate Cancer Study Collaborators; British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators, Horwich, A., Huddart, R.A., Khoo, V.S., Parker, C.C., Woodhouse, C.J., Thompson, A., Christmas, T., Ogden, C., Fisher, C., Jamieson, C., Cooper, C.S., English, D.R., Hopper, J.L., Neal, D.E., Easton, D.F. (2008). Multiple newly identified loci associated with prostate cancer susceptibility. Nature Genet 40, 316-321.

[18] Elliott, K.S., Zeggini, E., McCarthy, M.I., Gudmundsson, J., Sulem, P., Stacey, S.N. et al. (2010). Evaluation of association of HNF1B variants with diverse cancers: collaborative analysis of data from 19 genome-wide association studies. PLoS One. 2010; 5: e10858.

[19] George, B., Seals, S., Aban, I. (2014). Survival analysis and regression models. J Nucl Cardiol 21(4), 686-94.

[20] Ghadimi, M., Mahmoodi, M., Mohammad, K., Zeraati, H., Rasouli, M., Sheikhfathollahi, M. (2011). Family history of the cancer on the survival of the patients with gastrointestinal cancer in northern Iran, using frailty models. BMC Gastroenterol 11,104.

[21] Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J.T., Thorleifsson, G., Manolescu, A. et al. (2007). Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet 39, 977–983.

[22] Hegele, R.A., Cao, H., Harris, S.B., Zinman, B., Hanley, A.J., Anderson, C.M. (2000). Gender, obesity, hepatic nuclear factor-1alpha G319S and the age-of-onset of type 2 diabetes in Canadian Oji-Cree. Int J Obes Relat Metab Disord 24(8),1062-4.

[23] Hernán MA, Cole SR, Margolick J, Cohen M, Robins JM (2005) Structural accelerated failure time models for survival analysis in studies with time-varying treatments. Pharmacoepidemiol Drug Saf 14(7):477-91.

[24] Hjelmborg, J.B., Scheike, T., Holst. K., Skytthe, A., Penney, K.L., Graff, R.E., Pukkala, E., Christensen, K., Adami, H.O., Holm, N.V., Nuttall, E., Hansen, S., Hartman, M., Czene, K., Harris, J.R., Kaprio, J., Mucci, L.A. (2014). The heritability of prostate cancer in the Nordic Twin Study of Cancer. Cancer Epidemiol Biomarkers Prev 23(11), 2303-10.

[25] Jones, J.S., Amos, C.I., Pande, M., Gu, X., Chen, J., Campos, I.M., Wei, Q., Rodriguez-Bigas, M., Lynch, P.M., Frazier, M.L. (2006). DNMT3b polymorphism and hereditary nonpolyposis colorectal cancer age of onset. Cancer Epidemiol Biomarkers Prev 15(5), 886-91.

[26] Jones, J.S., Chi, X., Gu, X., Lynch, P.M., Amos, C.I., Frazier, M.L. (2004). p53 polymorphism and age of onset of hereditary nonpolyposis colorectal cancer in a Caucasian population. Clin Cancer Res 10(17), 5845-9.

[27] Kasza J, Wraith D, Lamb K, Wolfe R (2014) Survival analysis of time-to-event data in respiratory health research studies. Respirology 19(4):483-92.

[28] Kim, L., Liao, J., Zhang, M., Talamonti, M., Bentrem, D., Rao, S., Yang, G.Y. (2008). Clear cell carcinoma of the pancreas: histopathologic features and a unique biomarker: hepatocyte nuclear factor-1beta. Mod Pathol 21,1075–83.

[29] Klein, J.P., Moeschberger, M.L. (2003). Survival analysis: Techniques for censored and truncated data. New York: Springer.

[30] Köhler, H.F., Kowalski, L.P. (2012). A critical appraisal of different survival techniques in oral cancer patients. Eur Arch Otorhinolaryngol 269(1), 295-301.

[31] Krüger, S., Bier, A., Engel, C., Mangold, E., Pagenstecher, C., von Knebel Doeberitz, M., Holinski-Feder, E., Moeslein, G., Schulmann, K., Plaschke, J., Rüschoff, J., Schackert, H.K.; German Hereditary Non-Polyposis Colorectal Cancer Consortium. (2005). The p53 codon 72 variation is associated with the age of onset of hereditary non-polyposis colorectal cancer (HNPCC). J Med Genet 42(10), 769-73.

[32] Kulminski, A.M., Culminskaya, I., Ukraintseva, S.V., Arbeev, K.G., Arbeeva, L., Wu, D., Akushevich, I., Land, K.C., Yashin, A.I. (2011). Trade-off in the effects of the apolipoprotein E polymorphism on the ages at onset of CVD and cancer influences human lifespan. Aging Cell 10(3), 533-41.

[33] Levin, A.M., Machiela, M.J., Zuhlke, K.A., Ray, A.M., Cooney, K.A., Douglas, J.A. (2008). Chromosome 17q12 variants contribute to risk of early-onset prostate cancer. Cancer Res 68, 6492-6495.

[34] Li, X., Li, L., Fang, R. (2016).Statistical analysis of survival times based on proportional generalized odds models. J Data Sci, 14, 571-584.

[35] Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343(2), 78-85.

[36] Locatelli, I., Lichtenstein, P., Yashin, A.I. (2004). The heritability of breast cancer: a Bayesian correlated frailty model applied to Swedish twins data. Twin Res 7(2), 182-91.

[37] McCarty, C.A., Peissig, P., Caldwell, M.D., Wilke, R.A. (2008). The Marshfield Clinic Personalized Medicine Research Project: 2008 scientific update and lessons learned in the first 6 years. Personalized Medicine 5, 529-542.

[38] McCarty, C.A., Wilke, R.A., Giampietro, P.F., Wesbrook, S.D., Caldwell, M.D. (2005). Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Personalized Medicine 2, 49-79.

[39] Moghimi-Dehkordi, B., Safaee, A., Pourhoseingholi, M.A., Fatemi, R., Tabeie, Z., Zali, M.R. (2008). Statistical comparison of survival models for analysis of cancer data. Asian Pac J Cancer Prev 9(3), 417-20.

[40] Mustafa, A., El-Desouky, B.S., AL-Garash, S. (2016). The Weibull generalized flexible Weibull extension distribution. J Data Sci, 14, 453-478.

[41] Pankratz, V.S., de Andrade, M., Therneau, T.M. (2005). Random-effects Cox proportional hazards model: general variance components methods for time-to-event data. Genet Epidemiol 28(2), 97-109.

[42] Pettitt, A. N. (1984). Proportional odds models for survival data and estimates using ranks. Applied Statistics 33, 169-175.

[43] Pierce, B.L., Ahsan, H. (2010). Genetic susceptibility to type 2 diabetes is associated with reduced prostate cancer risk. Hum Hered. 69, 193-201.

[44] Pourhoseingholi, M.A., Hajizadeh, E., Moghimi Dehkordi, B., Safaee, A., Abadi, A., Zali, M.R.(2007). Comparing Cox regression and parametric models for survival of patients with gastric carcinoma. Asian Pac J Cancer Prev 8(3), 412-6.

[45] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38, 904-9.

[46] Pu, S., Oluyede, B.O., Qiu, Y., Linder. D. (2016). A generalized class of exponentiated modified Weibull distribution with applications. J Data Sci, 14, 453-478.

[47] Rebouissou, S., Vasiliu, V., Thomas, C., Bellanne-Chantelot, C., Bui, H., Chretien, Y., Timsit, J., Rosty, C., Laurent-Puig, P., Chauveau, D., Zucman-Rossi, J. (2005). Germline hepatocyte nuclear factor 1-alpha and 1-beta mutations in renal cell carcinomas. Hum Molec Genet 14, 603-614.

[48] Reeves, S.G., Meldrum, C., Groombridge, C., Spigelman, A.D., Suchy, J., Kurzawski, G., Lubinski, J., McElduff, P., Scott, R.J. (2009). MTHFR 677 C>T and 1298 A>C polymorphisms and the age of onset of colorectal cancer in hereditary nonpolyposis colorectal cancer. Eur J Hum Genet 17(5), 629-35.

[49] Setiawan, V.W., Haessler, J., Schumacher, F., Cote, M.L., Deelman, E., Fesinmeyer, M.D., Henderson, B.E., Jackson, R.D., Vöckler, J.S., Wilkens, L.R., Yasmeen, S., Haiman, C.A., Peters, U., Le Marchand, L., Kooperberg, C. (2012). HNF1B and endometrial cancer risk: results from the PAGE study. PLoS One 7,e30390.

[50] Shugart, Y.Y., Hemminki, K., Vaittinen, P., Kingman, A., Dong, C.(2000). A genetic study of Hodgkin's lymphoma: an estimate of heritability and anticipation based on the familial cancer database in Sweden. Hum Genet 106(5), 553-6.

[51] Simonoff, J.S. (2003). Analyzing Categorical Data, New York: Springer-Verlag.

[52] Sotamaa, K., Liyanarachchi, S., Mecklin, J.P., Järvinen, H., Aaltonen, L.A., Peltomäki, P., de la Chapelle, A. (2005). p53 codon 72 and MDM2 SNP309 polymorphisms and age of colorectal cancer onset in Lynch syndrome. Clin Cancer Res 11(19 Pt 1), 6840-4.

[53] Spurdle, A.B., Thompson, D.J., Ahmed, S., Ferguson, K., Healey, C.S., O'Mara, T. et al. (2011). Genome-wide association study identifies a common variant associated with risk of endometrial cancer. Nat Genet 43, 451–454.

[54] Sun, J.Z., Yang, X.X., Hu, N.Y., Li, X., Li, F.X., Li, M. (2011). Genetic Variants in MMP9 and TCF2 Contribute to Susceptibility to Lung Cancer. Chin J Cancer Res 23, 183-7.

[55] Terasawa, K., Toyota, M., Sagae, S., Ogi, K., Suzuki, H., Sonoda, T., Akino, K., Maruyama, R., Nishikawa, N., Imai, K., Shinomura, Y., Saito, T., Tokino, T. (2006). Epigenetic inactivation of TCF2 in ovarian cancer and various cancer cell lines. Br J Cancer 94, 914-21.

[56] Thomas, G., Jacobs, K.B., Yeager, M., Kraft, P., Wacholder, S., Orr, N. et al. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet 40, 310-315.

[57] Wang, K.S., Liu, X., Zheng, S., Zeng, M., Pan, Y., Callahan, K. (2012). A novel locus for body mass index on 5p15.2: a meta-analysis of two genome-wide association studies. Gene 500(1), 80-84.

[58] Wang, K.S., Owusu, D., Pan, Y., Xu, C. (2014). Common genetic variants in the HNF1B gene contribute to type 2 diabetes and multiple cancers. Austin Biomarkers and Diagnosis 1(1):id1002.

[59] Wang, K.S., Pan, Y., Wang, W., Xu, C. (2015). Bayesian survival analysis of genetic variants in PTPRN2 gene for age at onset of cancer. International Journal of Clinical Biostatistics and Biometrics 1(1):004.

[60] Wang, S.J., Kalpathy-Cramer, J., Kim, J.S., Fuller, C.D., Thomas, C.R.(2010). Parametric survival models for predicting the benefit of adjuvant chemoradiotherapy in gallbladder cancer. AMIA Annu Symp Proc. 10, 847-51.

[61] Waters, K.M., Henderson, B.E., Stram, D.O., Wan, P., Kolonel, L.N., Haiman, C.A. (2009). Association of diabetes with prostate cancer risk in the multiethnic cohort. Am J Epidemiol 169, 937–945.

[62] Zhang, Y.R., Xu, Y., Yang, K., Liu, M., Wei, D., Zhang, Y.G., Shi, X.H., Wang, J.Y., Yang, F., Wang, X., Liang, S.Y., Zhao, C.X., Wang, F., Chen, X., Sun, L., Zhu, X.Q., Zhu, L., Yang,Y.G., Tang, L., Jiao, H.Y., Huo, Z.H., Yang, Z. (2012). Association of six susceptibility Loci with prostate cancer in northern Chinese men. Asian Pac J Cancer Prev 13, 6273-6.

[63] Zhu, H.P., Xia, X., Yu, C.H., Adnan, A., Liu, S.F., Du, Y.K. (2011). Application of Weibull model for survival of patients with gastric cancer. BMC Gastroenterol 11,1.

Table 1: Descriptive Characteristics of Cases and Controls in the study sample

| | Non-Cancer | Cancer |
|---|---|---|
| Number | 2848 | 716 |
| Sex, N (%) | | |
| Males | 1135(40%) | 340(47%) |
| Females | 1713(60%) | 376(53%) |
| Obesity, N (%) | | |
| No | 1700(60%) | 422(59%) |
| Yes | 1148(40%) | 294(41%) |
| Alcohol use, N (%) | | |
| No | 1060(37%) | 288(40%) |
| Yes | 1783(63%) | 425(60%) |
| Smoking status, N (%) | | |
| Never | 1487(52%) | 327(46%) |
| Current | 254(9%) | 54(7%) |
| Past | 1104(39%) | 331(47%) |
| Age, years | | |
| Mean ± SD | 65.1±11.3 | 71.1±10.3 |
| Range | 46-90 | 46-90 |
| Age at onset, years | | |
| Mean ± SD | - | 64.2±12.8 |
| Range | - | 23-90 |

Table 2: Results of the Cox Regression and Parametric Models in the Multivariate Survival Analysis of AAO of Cancer

| Models | AIC[a] | BIC[b] | AIC[c] | BIC[d] | AIC[e] | BIC[f] | AIC[g] | BIC[h] | AIC[i] | BIC[j] |
|---|---|---|---|---|---|---|---|---|---|---|
| Cox | 7955.52 | 7987.47 | 7952.80 | 7984.75 | 7917.22 | 7949.15 | 7943.91 | 7975.87 | 7935.73 | 7967.67 |
| Weibull | 5594.59 | 5635.67 | 5590.97 | 5632.06 | 5569.69 | 5610.73 | 5582.38 | 5623.38 | 5578.04 | 5619.11 |
| Exponential | 7344.54 | 7381.07 | 7344.32 | 7380.84 | 7314.77 | 7351.26 | 7343.91 | 7380.43 | 7343.91 | 7380.43 |
| Log-logistic | 5685.02 | 5726.11 | 5678.06 | 5719.15 | 5654.21 | 5695.25 | 5665.00 | 5706.09 | 5661.89 | 5702.97 |
| Log-normal | 5705.60 | 5746.69 | 5700.06 | 5741.15 | 5676.62 | 5717.67 | 5690.32 | 5731.40 | 5686.24 | 5727.32 |
| Gamma | 5595.50 | 5641.16 | 5592.22 | 5637.88 | 5571.02 | 5616.63 | 5583.94 | 5629.60 | 5579.54 | 5625.18 |

[a,b] AIC and BIC for rs3110649 adjusted for sex, alcohol use, smoking status, and obesity; [c,d] AIC and BIC for rs757210 adjusted for sex, alcohol use, smoking status, and obesity; [e,f] AIC and BIC for rs430796 adjusted for sex, alcohol use, smoking status, and obesity; [g,h] AIC and BIC for rs4239217 adjusted for sex, alcohol use, smoking status, and obesity; [i,j] AIC and BIC for rs7501939 adjusted for sex, alcohol use, smoking status, and obesity.

Table 3: Survival Analysis of the 5 SNPs Associated with AAO Using the Weibull Model

| SNP | Position (bp) | Allele[a] | MAF[b] | HWE[c] | Genotype | β[d] | 95%CI[e] | p-value[f] | TR[g] | 95%CI[h] |
|---|---|---|---|---|---|---|---|---|---|---|
| rs3110649 | 33144293 | T | 0.21 | 0.374 | | | | | | |
| | | | | | CC | 0.063 | 0.003,0.12 | 0.0387 | 1.07 | 1.01,1.13 |
| | | | | | CT | 0.045 | -0.02,-0.10 | 0.143 | 1.05 | 0.98,1.11 |
| | | | | | TT | | | | 1.00 | |
| rs757210 | 33170628 | A | 0.37 | 0.613 | | | | | | |
| | | | | | AA | -0.056 | -0.097,-0.019 | 0.0039 | 0.93 | 0.89,0.97 |
| | | | | | AG | -0.001 | -0.028,0.025 | 0.933 | 1.00 | 0.97,1.03 |
| | | | | | GG | | | | 1.00 | |
| rs430796 | 33172153 | G | 0.48 | 0.129 | | | | | | |
| | | | | | AA | 0.034 | -0.001,0.069 | 0.0567 | 1.03 | 1.00,1.07 |
| | | | | | AG | 0.034 | 0.002,0.066 | 0.0372 | 1.03 | 1.01,1.07 |
| | | | | | GG | | | | 1.00 | |
| rs4239217 | 33173100 | G | 0.4 | 0.18 | | | | | | |
| | | | | | AA | 0.075 | 0.038,0.111 | <0.0001 | 1.08 | 1.04,1.12 |
| | | | | | AG | 0.076 | 0.039,0.111 | <0.0001 | 1.08 | 1.04,1.12 |
| | | | | | GG | | | | 1.00 | |
| rs7501939 | 33175269 | T | 0.39 | 0.073 | | | | | | |
| | | | | | CC | 0.069 | 0.031,0.106 | 0.0004 | 1.07 | 1.03,1.11 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | CT | 0.069 | 0.033,0.106 | 0.0002 | 1.07 | 1.03,1.11 |
| | TT | | | | 1.00 | |

[a] Minor allele; [b] Minor allele frequency; [c] Hardy-Weinberg equilibrium test p-value; [d] Regression coefficient for AAO of cancer based on the Weibull model; [e] 95%CI of regression coefficient for AAO of cancer based on the Weibull model; [f] p-value for AAO of cancer based on Weibull model; [g] Time ratio (TR) for the genotype comparing with reference; [h] 95%CI of TR for the genotype comparing with reference .

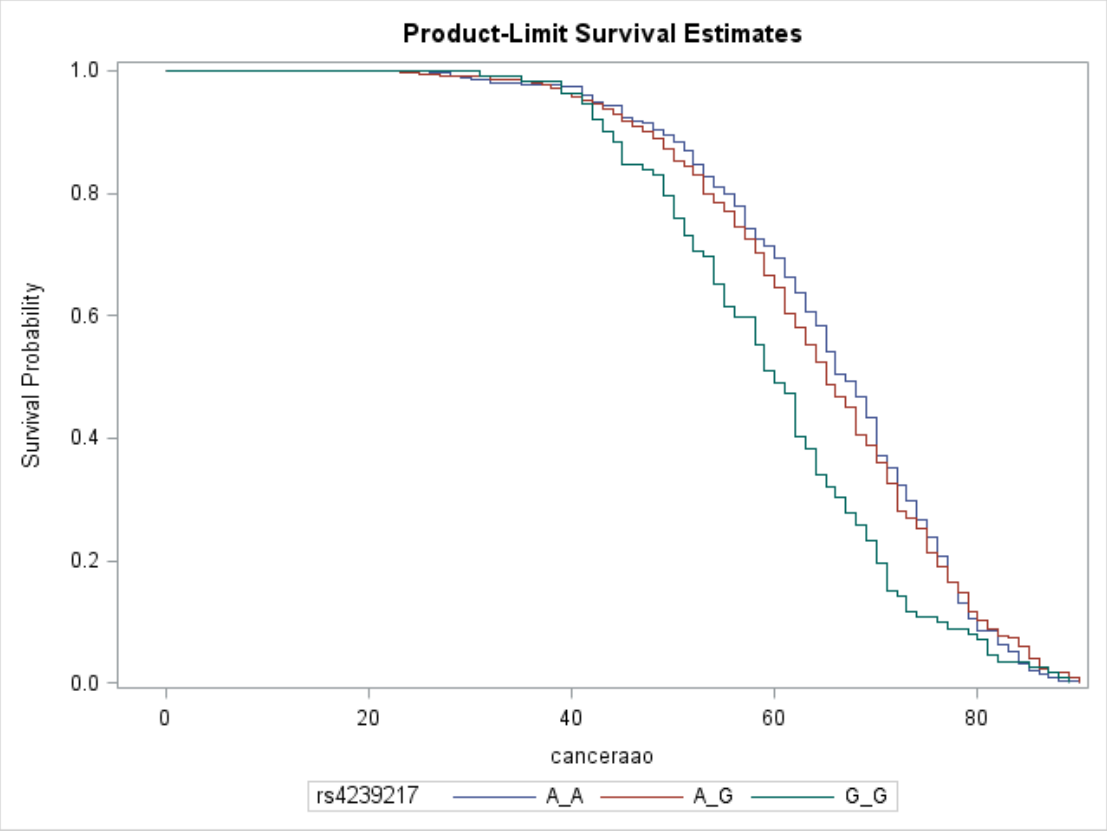


Figure 1: Survival Function by Three Genotypes of rs4239217

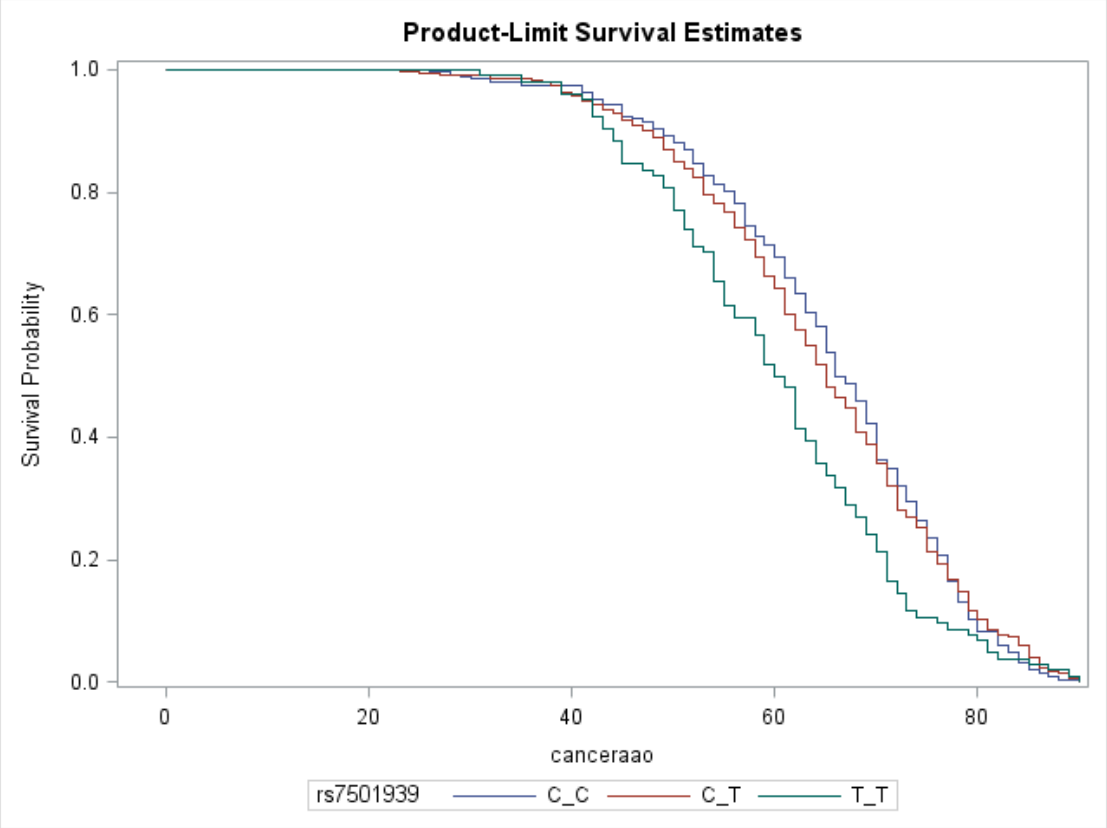


Figure 2: Survival Function by Three Genotypes of rs7501939

## APPENDIX

The following program using PROC PHREG showed one SNP rs4239217, sex, alcohol use, smoking status, and obesity with the AAO of cancer. The rs4239217 has 3 genotypes - A_A, A_G and G_G, respectively; while the G_G genotype was considered as the reference.

**proc phreg** data= canceraao;

class sex(ref="1") obesity(ref="1") alcohol(ref="1") smoking(ref="1") rs4239217 (ref = "G_G")/param=ref;

model canceraao*statuscan(**0**) = sex obesity alcohol smoking rs4239217 / risklimits ;

hazardratio rs4239217;

**run**;

The following program using PROC LIFEREG showed one SNP rs4239217, sex, alcohol use, smoking status, and obesity with the AAO of cancer for the Weibull distribution. Other parametric models can be tested by changing the dist (distribution) option in the model.

**proc lifereg** data= canceraao;

class sex obesity  alcohol smoking rs4239217;

model canceraao*statuscan(**0**) = sex obesity alcohol smoking rs4239217 / dist=WEIBULL;

**run**;

Kesheng Wang

Department of Biostatistics and Epidemiology

College of Public Health

East Tennessee State University, Johnson City, TN 37614, USA

wangk@etsu.edu

Xuefeng Liu

Department of Systems Leadership and Effectiveness Science

School of Nursing

University of Michigan, Ann Arbor, MI 48109-5482, USA

Yue Pan

Department of Public Health Sciences

Miller School of Medicine

University of Miami, Miami, FL 33136, USA

Daniel Owusu

Department of Biostatistics and Epidemiology

College of Public Health

East Tennessee State University, Johnson City, TN 37614, USA

Chun Xu

Department of Health and Biomedical Sciences

College of Health Affairs

University of Texas Rio Grande Valley, Brownsville, TX 78520, USA