Global Journal of Medical Research: F
Diseases
Volume 21 Issue 3 Version 1.0 Year 2021
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4618 & Print ISSN: 0975-5888

Survival Analysis for Pancreatic Cancer Patients using Cox-Proportional Hazard (CPH) Model

By Aditya Chakraborty & Chris P. Tsokos

Abstract- Pancreatic cancer is comparatively rare but extremely lethal. In the United States, pancreatic cancer is the 4th leading cause of cancer death, and in Europe, it is the 6th. Though Pancreatic cancer remains incurable if detected late, research into improving the therapeutic strategy has increased significantly in recent years. However, it is ambiguous if sustained improvements have been achieved by identifying the most prominent risk factors responsible for cancer. In this article, we studied the survival times of 677 pancreatic cancer patients with *fifteen* risk factors. The semi-parametric Cox proportional hazard (CPH) model was used to examine the covariate effect taking into account all of the statistically significant risk factors and their significant twoway interactions. A careful and rigorous assessment of the risk factors based on the AIC of the stepwise selection technique revealed seven risk factors, and ten interaction terms are statistically significantly contributing to the survival times. The final Cox-PH model was well-validated and satisfied all the key assumptions. The identified risk factors and their interactions are ranked according to the prognostic effect on the survival time based on the hazard ratio. We found the most contributing risk factor is the combined effect of patients with emphysema and cancer stage regional with a hazard ratio (HR) = 8.84.

Keywords: pancreatic cancer, cox-PH model, pancreatic survival function.

GJMR-F Classification: NLMC Code: WI 800

Survival Analysis for Pancreatic Cancer Patients using Cox-Proportional Hazard (CPH) Model

Aditya Chakraborty [α] & Chris P. Tsokos [σ]

Abstract- Pancreatic cancer is comparatively rare but extremely lethal. In the United States, pancreatic cancer is the 4*th* leading cause of cancer death, and in Europe, it is the 6*th*. Though Pancreatic cancer remains incurable if detected late, research into improving the therapeutic strategy has increased significantly in recent years. However, it is ambiguous if sustained improvements have been achieved by identifying the most prominent risk factors responsible for cancer. In this article, we studied the survival times of 677 pancreatic cancer patients with *fifteen* risk factors. The semi-parametric Cox proportional hazard (CPH) model was used to examine the covariate effect taking into account all of the statistically significant risk factors and their significant twoway interactions. A careful and rigorous assessment of the risk factors based on the AIC of the stepwise selection technique revealed seven risk factors, and ten interaction terms are statistically significantly contributing to the survival times. The final Cox-PH model was well-validated and satisfied all the key assumptions. The identified risk factors and their interactions are ranked according to the prognostic effect on the survival time based on the hazard ratio. We found the most contributing risk factor is the combined effect of patients with emphysema and cancer stage regional with a hazard ratio (HR) = 8.84. The most significant highest contributing individual risk factor is diabetes with a hazard ratio of 2.39, followed by ibuprofen with a hazard ratio of 1.83. This study offers prognostic and therapeutic significance for further enhancement in the treatment strategy of pancreatic cancer.

Keywords: pancreatic cancer, cox-PH model, pancreatic survival function.

I. Introduction

In the domain of the lethal carcinogenic diseases affecting humans, pancreatic cancer is one of the fatal cancers and continues to be a crucial unsolved health problem at the start of the 21st century. Because of the high fatality rates, pancreatic cancer incidence rates are almost equal to mortality rates (22). According to the current health science researchers, this disease causes approximately 30,000 deaths per year in the USA.(1). It is the fourth principal reason for cancer death in the USA and leads to an estimated 227,000 deaths per year worldwide. The incidence and number of deaths caused by pancreatic tumors have been gradually increasing, even as incidence and mortality of other common cancers have been declining. Despite developments in detection and management of pancreatic cancer, only about 4% of patients will live five years after diagnosis, (2). The normal pancreas consists of digestive enzyme-secreting acinar cells, bicarbonate-secreting ductal cells, centroacinar cells that are the geographical transition between acinar and ductal cells, hormone-secreting endocrine islets and relatively inactive stellate cells. The majority of malignant neoplasms of the pancreas are adenocarcinomas. Rare pancreatic neoplasms include neuroendocrine tumors (which can secrete hormones such as insulin or glucagon) and acinar carcinomas (which can release digestive enzymes into the circulation). Particularly, ductal adenocarcinoma is the most frequent kind of malignancy of the pancreas; this tumor (commonly referred to as pancreatic cancer) presents a substantial health problem, with an estimated 367,000 new cases diagnosed worldwide in 2015 and an associated 359,000 deaths in the same year(3)(4). After the detection of pancreatic cancer, doctors usually perform some additional tests to understand better if cancer has been spread or the spreading area of cancer. Different imaging tests, such as a PET scan, can help doctors identify the presence of cancerous growths. With these tests, doctors try to establish cancer's stage. Staging helps explicate how advanced the cancer is. It also assists doctors in deciding the treatment options. The following are the description of the stages used in our dataset according to the definition of the Surveillance, Epidemiology, and End Results (SEER) database.

Author α σ: e-mails: adityachakra@usf.edu, ctsokos@usf.edu

1. **Localized**: There is no sign that the cancer has spread outside of the pancreas.
2. **Regional**: The cancer has spread from the pancreas to nearby structures or lymph nodes.
3. **Distant**: The cancer has spread to distant parts of the body such as the lungs, liver or bones.

The following Figure 1 shows the different parts of the pancreas.

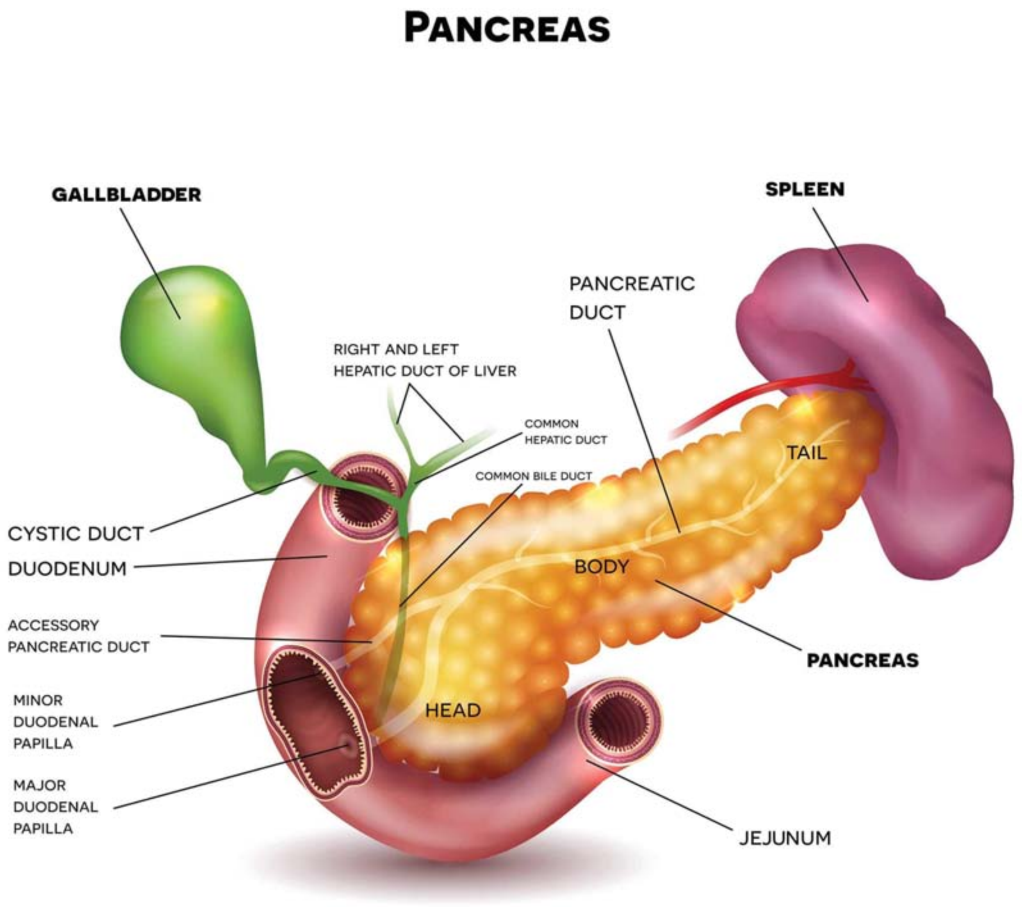

Figure 1: Different Parts of the Pancreas

Although, in most cases, pancreatic cancer remains incurable, researchers have focused on how to improve the survival times of patients diagnosed with pancreatic cancer. Cox proportional hazard model/ Cox model (5) has been used extensively in the literature of cancer research to address the hazard of an individual patient with respect to specific risk factors. It is also useful to assess the association between different treatments and the survival time of patients. Perera and Tsokos (6) developed a statistical model with Non-Linear Effects and Non-Proportional Hazards for Breast Cancer Survival Analysis. In their study, the authors have identified the effects of age and breast cancer tumor size at diagnosis on the hazard function, which have a non-linear effect. Also, they have addressed the different assumptions of the proportional hazard model. Asano, Hirakawa, and Hamada (7) used an imputation-based receiver operating characteristic curve (AUC) to evaluate the predictive accuracy of the cure rate from the PH cure model. They also illustrated the estimation of the imputation-based AUCs using breast cancer data. Yong & Tsokos (8) have evaluated the effectiveness of widely used Kaplan-Meier (KM) model, non-parametric Kernel density (KD) models with the Cox PH model, using both Monte Carlo simulations on the breast cancer data. Du, Li et al. (2018) (9) compared a flexible parametric survival model (FPSM) and Cox model using Markov transition probabilities from a cohort study data investigating ischemic stroke outcomes in Western China. The FPSM produced hazard ratio and baseline cumulative hazard estimates similar to those obtained using the Cox proportional hazards model. Mamudu & Tsokos (20) developed a semi-parametric Cox model for Multiple Myeloma Cancer (MMC) patients and addressed the validity of the assumptions of the model.

In our study, we used the semi-parametric Cox-PH survival analysis of the survival times to estimate the survival rate of patients diagnosed with pancreatic cancer. We utilized the Cox-PH model to analyze the proportion of survival time, taking into account the fifteen risk factors that are identified in section 2.1. We assessed the relationship between the proportion of survival time as a function of the attributable risk factors and two-way interactions based on the Cox proportional hazard (PH) model. The significant attributable risk factors identified were meticulously investigated and selected based on the step-wise model selection method, with the final model representing the model with the least AIC. The final Cox-PH model was validated to satisfy all the main assumptions of the Cox-PH model.

II. Methodology

a) Data Description

The data for our study has been obtained from The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial system of the National Cancer Institute (NIH) database. The data contains information on patients diagnosed with pancreatic adenocarcinoma. We are concerned with the survival time (in days) and cause-specific death (deaths due to pancreatic cancer) for each patient. The survival time of patients is one of the most important factors used in all cancer research. It is important to evaluate the severity of cancer, which helps to decide the prognosis and help identify the correct treatment methods. There were a total of 677 patient information in our study after eliminating the missing observations for which several risk factors were missing. In our study, the response variable is the survival time of patients (in days). There are a total of *fifteen* risk factors used in our survival model. Twelve of them are categorical, and three of them are numeric variables. The description of the risk factors is as follows.

1. Age (Numeric) (X_1): Age of diagnosis of the patient.
2. Stage (Categorical) (X_2): Pancreatic Cancer Stages, categorized as a) localized, b) regional, and c) distant
3. Aspirin (Categorical) (X_3): Does the person use Aspirin Regularly?
4. Ibuprofen (Categorical) (X_4): Does the person use Ibuprofen Regularly?
5. Relatives (Categorical) (X_5): The number of first-degree relatives with pancreatic cancer.
6. Diabetes (Categorical) (X_6): Did the patient ever have diabetes?
7. Heart attack (Categorical) (X_7): Did the participant ever have coronary heart disease or a heart attack?
8. Emphysema (Categorical) (X_8): Did the patient ever have emphysema?
9. Sex (Categorical) (X_9): Sex of the individual.
10. BMI (numeric) (X_{10}): Current Body Mass Index (BMI) at Baseline (In lb/in2)
11. Cigarette Years (numeric) (X_{11}) : The total number of years the patient smoked.
12. Diverticulosis (Categorical) (X_{12}): Did the participant ever have diverticulitis or diverticulosis?
13. Smoke (Categorical) (X_{13}): Has the patient ever smoked cigarettes regularly for six months or longer?

14. Gallbladder (Categorical) (X_{14}): Did the individual ever have gall bladder stones or inflammation?
15. Hypertension (Categorical) (X_{15}): Did the individual ever have high blood pressure?

A schematic diagram of the data used in our study with the description of risk factors is shown in Figure 2, below.

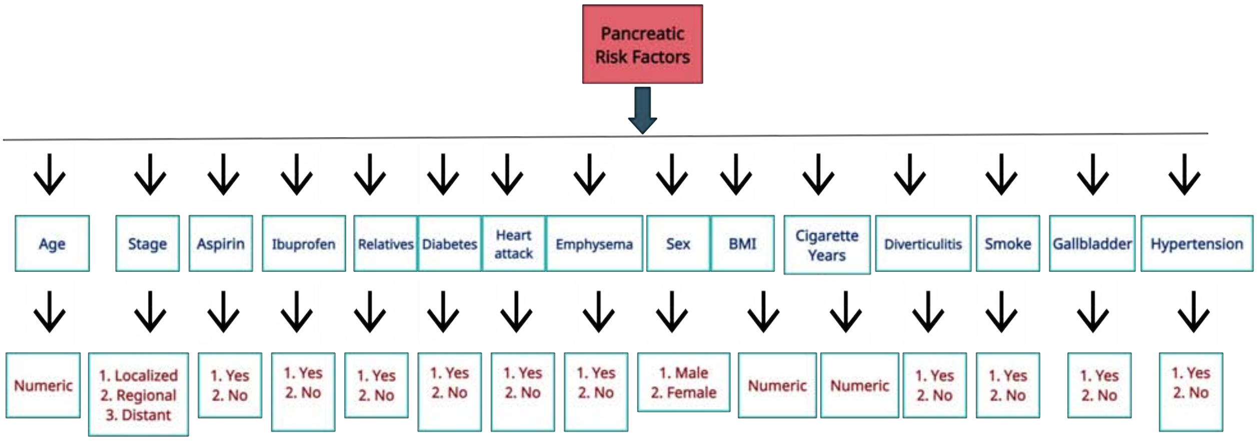

Figure 2: Pancreatic Cancer Data with Relevant Risk Factors

As the above Figure illustrates, we see that twelve out of fifteen risk factors are categorical, having two or more categories. Before we proceed with our main analysis, it is very important to investigate if there is any statistically significant difference between the survival times of male and female patients diagnosed with pancreatic cancer. If any significant differences are found, separate analyses for each gender should be performed. To answer this question, we used the non-parametric Wilcoxon rank-sum test with continuity correction and obtained a p-value of .47, indicating that there is not enough sample evidence to reject the following null hypothesis (H_0) at a 5% level of significance.

H_0: There is no statistically significant difference between the survival times of male and female patients.

Thus we proceeded with our analysis and modeling by combining the male and female data together to constitute our sample size.

III. Brief Description of Cox Proportional Hazard (CPH) Model

The Cox PH model, proposed by Sir David Cox, is a statistical method that can be used for survival-time (time-to-event) outcomes on one or more risk factors and their interactions. In survival analysis, the Cox model has been widely recommended for semi-parametric modeling of the survival time relationship as a function of the risk factors. Kleinbaum & Klein (10) gives a good introductory review of the background and methodology, and more detailed descriptions have been provided by Kalbeisch , and Prentice (11)(12). In this section, we give a brief review of the Cox proportional hazards model. An important aspect of the Cox PH model is the hazard function $h(t)$. It measures the rate of the event of occurrence (death) as a function of time t. We define the hazard function as follows; Let random variable T denotes the survival time with cumulative density function $F_T(t)$, given by

$$F_T(t) = P(T \leq t) = \int_0^t f(t)dt \ ,$$

where $f(t) = \frac{dF_T(t)}{dt}$ is the probability density function (pdf) of the random variable T. The survival function at time t is defined as:

$$S(t) = P(T \geq t) = 1 - F_T(t) = \int_t^{\infty} f(t)dt \ . \tag{1}$$

$S(t)$ gives the probability that a specific individual would survive beyond time t. Since $S(t)$ is a probability, $0 \leq S(t) \leq 1$ and $S(0) = 1$, for $T \geq 0$ from (1) we have,

$$f(t) = \frac{dF_T(t)}{dt} = -\frac{dS(t)}{dt} \ . \tag{2}$$

For continuous survival data, the hazard function plays a very important role. It aims to quantify the *instantaneous risks* that an event will occur at time t. It is defined as the follows:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P\{t \leq T < t + \Delta t \mid T \geq t\}}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P\{t \leq T < t + \Delta t\}}{\Delta t} \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)} \ . \end{aligned} \tag{3}$$

Combining (2) and (3), we obtain,

$$h(t) = -\frac{d}{dt} log\{S(t)\} \ . \tag{4}$$

Integrating both sides of equation (4) gives an expression for the survival function $S(t)$ in terms of the hazard function $h(t)$. That is,

$$S(t) = exp\Big[- \int_0^t h(u)du \Big] \ . \tag{5}$$

Now, from (3) and (5) we can express the pdf $f(t)$ as a function of $S(t)$ and $h(t)$ given by,

$$f(t) = h(t)exp\Big[- \int_0^t h(u)du \Big] \ . \tag{6}$$

From (3) the cumulative hazard function $H(t)$ can be expressed as:

$$H(t) = \int_0^t h(u)du = -lnS(t) \ . \tag{7}$$

Now, suppose $X_i = (X_{i1}, X_{i1}, \ldots, X_{ip})$ are the realized values of the risk factor for the i^{th} subject. Then, the Cox PH model (not including time-dependent risk factors or non-proportional hazards) can be expressed in term of the hazard as:

$$h_i(t) = \lambda_0(t)exp\Big[\sum_{j=1}^{p} \beta_j X_{ij} + \sum_{j \neq k} \eta_{jk} X_{ij} X_{ik} \Big] \ , \ \ j,k = 1,2,\ldots,p. \tag{8}$$

In the above expression, λ_0 is called the *baseline hazard* which can be thought of as the hazard function for an individual for which all value of the risk factors are 0. β_j measures the impact of X_{ij} on $h_i(t)$. η_{jk} is the interaction coefficient between j^{th} and k^{th} risk factor of the i^{th} individual and measures the impact of $X_{ij}X_{ik}$ on $h_i(t)$. From (8), it is clear that the individual hazard is a function of the risk factors and their interactions and is connected through baseline hazard. From (8), we can write,

$$ln\left\{\frac{h_i(t)}{h_k(t)}\right\} = \left[\sum_{j=1}^{p}\beta_j X_{ij} + \sum_{j\neq k}\eta_{jk}X_{ij}X_{ik}\right], \; j \neq k \qquad (9)$$

From the above expression we see that the ratio of log hazard of the i^{th} and k^{th} individual is constant over time. Thus, the name *proportional* in the Cox PH model. We interpret the hazard ratio (HR) in the following ways:

1. HR = 1; implies that there is no hazard effect. Thus, the risk factors have no relationship with the event probability, thus, no influence on the length of survival.
2. HR > 1 (i.e. equivalently $\beta_i > 0$), implies an increase in hazard. That is, the risk factors have a positive association with the event probability, thus, a negative association with the length of survival (bad prognostic factor).
3. HR < 1 (i.e. equivalently $\beta_i < 0$), implies a decrease in hazard. That is, the risk factors are negatively associated with the probability of the event, thus, positively associated with the length of survival (good prognostic factor).

A detailed description of the hazard ratio have been provided in (14) (15).

IV. Statistical Data Analysis and Survival Modeling

We now proceed to develop our most parsimonious statistical model using Cox PH. We initially started by fitting the Cox-PH model to the survival times t as a function of all fifteen risk factors given in Figure 2 together with their two-way interactions. So, there were fifteen risk factors and $\binom{15}{2} = 105$ two-way interaction terms. We used a stepwise model selection procedure to select the best model with the minimum Akaike information criterion $(AIC = 2ln(L) + 2k$, where L is the value of the maximum likelihood function of the model and k represents the number of estimated model parameters)(13). AIC gives an estimation of the relative amount of information missing in the model; hence, the smaller the AIC value, the better the quality of the model. It also deals with the risk associated with overfitting or under-fitting the model. One of the most important assumptions of the Cox PH is proportionality. Initially, all of the risk factors and two-way interactions except *age* satisfied the assumption. The range of the variable age was [50-90). So, we divided the range into two categories, say [50,70), and [70,90). Now, we use *stratification* on the variable age. Stratification is one of the tools used by researchers when one of the risk factors does not satisfy the proportionality assumption. The stratification will produce hazard ratios for all other risk factors in the presence of two hazards intrinsic to the level of age. Since age violated the proportional hazards assumption, stratifying it will help meet the PH assumption and provide more valid estimates for all other risk factors. The stratified model allows the baseline hazard $\lambda_0(t)$ to vary between strata but controls the effect of the risk factors to be the same for each stratum. For each subject in strata $s, s = 1, 2$, we have from (8),

$$h_i(t) = \lambda_{0s}(t)exp\left[\sum_{j=1}^{p}\beta_j X_{ij} + \sum_{j\neq k}\eta_{jk}X_{ij}X_{ik}\right], \; j,k = 1,2,\ldots,p. \; (s = 1,2) \qquad (10)$$

However, it is not possible to get an estimate of the risk factor (age) separately after stratification. The following Figure 3 illustrates the survival curve for the two age groups.

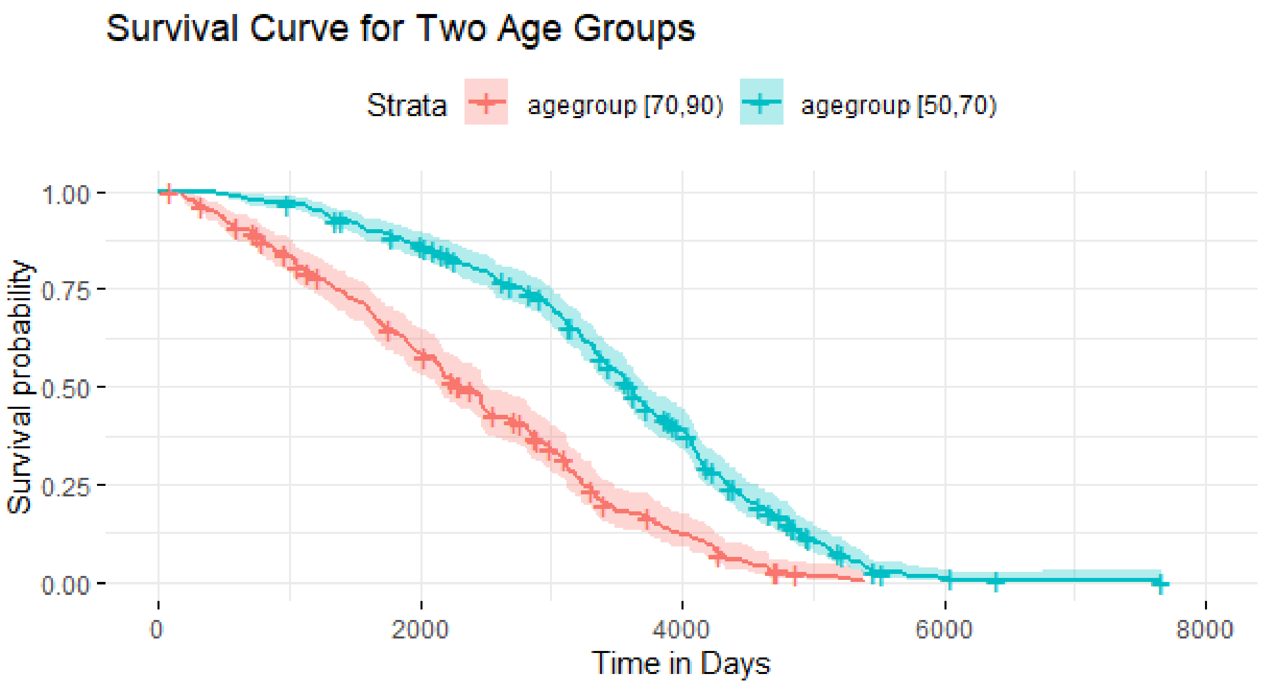

Figure 3: The Estimated Survival Curve for the two different Age Groups

We observe from Figure 3 that the age group [70,90) (highlighted in pink) is much more vulnerable than the age group [50,70) (highlighted in blue) in terms of survival probabilities. That is, a randomly selected patient in the age group [50,70) has a higher survival probability than a patient in the group [70,90), which is quite plausible.

The cumulative hazard function, $H(t)$, of the two age groups is given below by Figure 4.

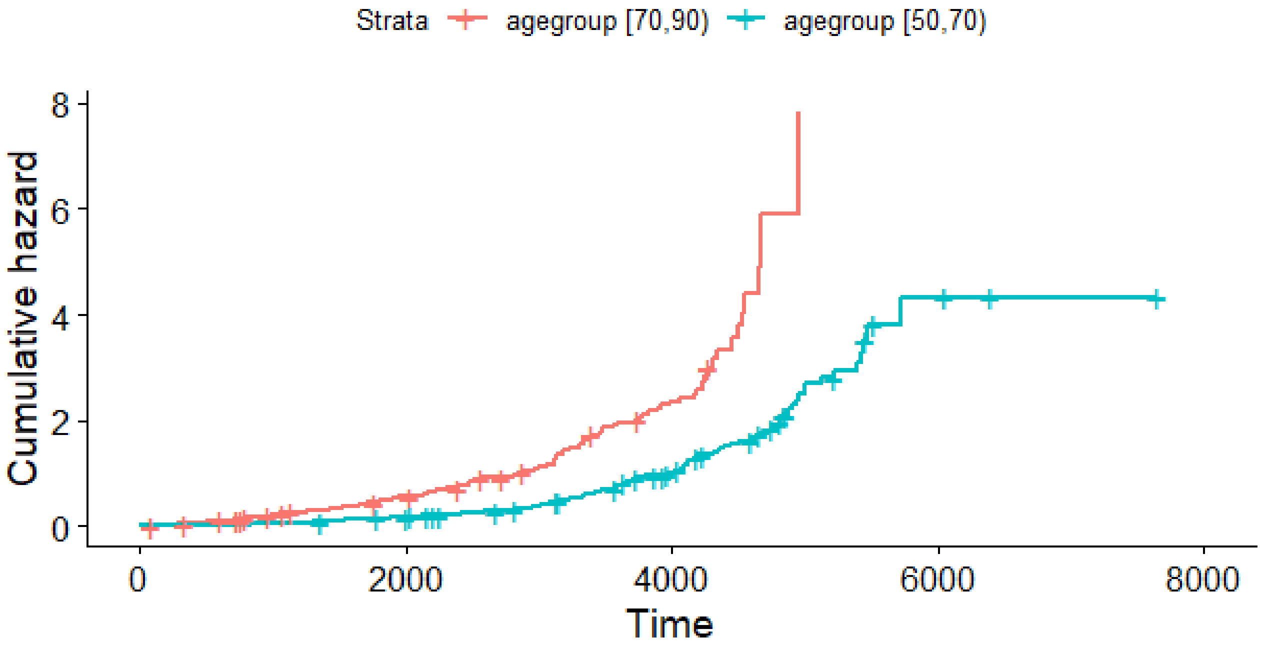

Figure 4: Cumulative Hazard Functions of the Two Age Groups

As the above figure 4 suggests, the cumulative hazard for patients in the age group [70,90) is more than patients belonging to [50,70). We see that the cumulative hazard is the same for two age groups, almost up to $t = 1000$ days. After that, the cumulative hazard is exponentially increasing for the age group [70,90). However, for the age group [50,70), the cumulative hazard has an increasing pattern up to $t = 5800$ days approximately. After that, the graph has a steady pattern.

The following Table 1 illustrates the count of each category of all the risk factors after stratification.

Table 1: Table Showing the Count of Different Categories of Risk Factors

Risk Factors		Count
Stage	Localized	135
	Regional	178
	Distant	364
Aspirin	Yes	333
	No	344
Ibuprofen	Yes	168
	No	509
Relatives	Yes	650
	No	27
Diabetes	Yes	83
	No	594
Heart attack	Yes	84
	No	593
Emphysema	Yes	19
	No	658
Sex	Male	388
	Female	289
BMI		677
Cigarette Years		677
Diverticulosis	Yes	41
	No	636
Smoke	Yes	404
	No	273
Gallbladder	Yes	98
	No	579
Hypertension	Yes	256
	No	421

The step-wise procedure produced *seven* out of fourteen significant risk factors and *ten* two-way interaction terms. There were some risk factors that did not contribute to the hazard individually, but, interacting with other risk factors, their effect was significant. Thus, we added those risk factors in our proposed model. That is why there are thirteen individual risk factors and ten interactions in the model (11). In the following model (11), we denote **"Y"** to indicate yes of a specific answer of a risk factor. That is, the specific category possesses the characteristic. For example, to answer the question "does the patient ever have diabetes?" the individual answers "yes." To describe any particular category of the risk factor *stage*, we use **L**, **R**, and **D** which are the first letters of Localized, Regional, and Distant. To describe male and female category of the variable *Sex*, we use the letters **M** and **F**, respectively. The most parsimonious model that we found after removing the insignificant (p-value > 0.05) term from the model is given as follows:

$$ln\widehat{\left[\frac{h_i(t)}{\lambda_0(t)}\right]} = \begin{cases} 0.3X_{2R} + .5X_{2D} - .53X_{3Y} \\ +.61X_{4Y} - .37X_{15Y} + .87X_{6Y} \\ -.6X_{5Y} - .7X_{8Y} \\ -.35X_{9F} + .0037X_{11} - .51X_{12Y} + .15X_{13Y} \\ +.28X_{14Y} - .56X_{4Y}X_{13Y} + .41X_{3Y}X_{9F} \\ +.6X_{3Y}X_{15Y} + .01X_{2R}X_{11} + .68X_{12Y}X_{9F} \\ +.32X_{15Y}X_{9F} - .47X_{15Y}X_{14Y} \\ -.52X_{2R}X_{4Y} + 2.18X_{2R}X_{8Y} + .8X_{15Y}X_{12Y} \end{cases} \quad (11)$$

Thus, the proposed statistical model consists of thirteen individual risk factors and ten interactions that contributes to the hazard.

a) *Estimating the Survival Function*

The above equation (10) can be written as:

$$h_i(t; X_{ij}, X_{ij}X_{ik}) = h_{0s}(t)exp\Big[\sum_{j=1}^{p}\hat{\beta}_j X_{ij} + \sum_{j\neq k}\hat{\eta}_{jk}X_{ij}X_{ik}\Big], \, j \neq k \quad (12)$$

We can express the Cox-PH model (11) in the form of the survival function, $S(t)$, by employing equation (5) from Section 3. Thus, the survival function of the Cox-PH model can be expressed as;

$$\begin{aligned} \hat{S}_i(t; X_{ij}, X_{ij}X_{ik}) &= exp\Big[-\int_0^t h_i(t; X_{ij}, X_{ij}X_{ik})dt\Big] \\ &= exp\Big[-\int_0^t h_{0s}(t)exp\Big[\sum_{j=1}^{p}\hat{\beta}_j X_{ij} + \sum_{j\neq k}\hat{\eta_{jk}}X_{ij}X_{ik}\Big]dt\Big] \\ &= exp\Big[exp\Big[\sum_{j=1}^{p}\hat{\beta}_j X_{ij} + \sum_{j\neq k}\hat{\eta_{jk}}X_{ij}X_{ik}\Big]\Big(-\int_0^t h_{0s}(t)dt\Big)\Big] \\ &= exp\Big(-\int_0^t h_{0s}(t)dt\Big)^{\left[\sum_{j=1}^{p}\hat{\beta}_j X_{ij}+\sum_{j\neq k}\hat{\eta_{jk}}X_{ij}X_{ik}\right]} \\ &= \big[S_{0s}(t)\big]^{\left[\sum_{j=1}^{p}\hat{\beta}_j X_{ij}+\sum_{j\neq k}\hat{\eta_{jk}}X_{ij}X_{ik}\right]} \end{aligned} \quad (13)$$

where $\hat{S_{is}}(t; X_{ij}, X_{ij}X_{ik})$ is the survival function at time t for i^{th} individual and $s^{th}, (s = 1, 2)$ stratum. $S_{0s}(t)$ is the baseline survivor function for each stratum $s = 1, 2$. After the estimation of $\hat{\beta}$ and $\hat{\eta}_{jk}$ by partial likelihood (16), $S_{0s}(t)$ can be estimated by a non-parametric maximum likelihood method (17). The co-efficient estimates of parameters $\hat{\beta}$ and $\hat{\eta}_{jk}$ are given in the third column of Table 2.

Table 2 below displays the estimates of the model coefficients/parameters, their hazard ratios (HR) $(exp(\hat{\beta}))$, standard error of coefficients, statistical significance, and 95% confidence interval. We proceed to rank the significant contributing risk factors and their significant interactions based on the prognostic effect on the survival times of patients diagnosed with pancreatic cancer using the hazard ratio (HR). Thus, we rank from the most contributing risk factor to the least contributing risk factor to pancreatic cancer patient's death or survival times.

Table 2: Ranking of the Significant Contributing Risk Factors and Interactions Based on Prognostic Effect to the Survival Time using the Hazard Ratios

Rank	Risk Factors	coeff($\hat{\beta}$)	HR [$exp(\hat{\beta})$]	[$S.E(\hat{\beta})$]	Lower 95%	Upper 95%
1	$X_{2R}X_{8Y}$	2.18	8.84	.96	1.32	59.1
2	X_{6Y}	.87	2.39	.33	1.2	4.6
3	$X_{15Y}X_{12Y}$	.8	2.28	.38	1.07	4.87
4	$X_{12Y}X_{9F}$	.68	1.98	.39	.92	4.25
5	X_{4Y}	.61	1.834	.25	1.27	2.62
6	$X_{3Y}X_{15Y}$	.6	1.831	.18	1.11	3.02
7	X_{2D}	.5	1.63	.17	1.16	2.3
8	$X_{3Y}X_{9F}$	.41	1.5	.18	1.06	2.13
9	$X_{15Y}X_{9F}$	.32	1.37	.18	.96	1.96
10	$X_{2R}X_{11}$	0.01	1.01	.007	.99	1.05
11	X_{9F}	-.35	.7	.13	.54	.91
12	X_{15Y}	-.37	.69	.16	.5	.95
13	$X_{15Y}X_{14Y}$	-.47	.63	.26	.42	.94
14	$X_{13Y}X_{4Y}$	-.46	.63	.2	.42	.94
15	X_{3Y}	-.53	.6	.13	.45	.77
16	$X_{2R}X_{4Y}$	-.52	.59	.3	.33	1.05
17	X_{5Y}	-.6	.55	.2	.35	.84

The above Table 2 describes different information, including the hazard ratio of all *seven* significant risk factors and all *ten* significant interactions used in the model. A positive estimated coefficient/weight ($\hat{\beta} > 0$) implies higher hazard rate, and thus a bad prognostic factor. on the contrary , a negative estimated coefficient/weight ($\hat{\beta} < 0$) implies a lower hazard rate, and thus a good prognostic factor. For example, $\hat{\beta_{9F}} = -0.35$ from Table 2, implies females are good prognostic of the survival time of pancreatic cancer; thus, females have a lower risk of death (higher survival rates) of cancer than males. The $exp(\hat{\beta})$ is the hazard ratio (HR). Thus, $exp(-0.35) = .7 < 1$ for gender female means being a female has a reduced risk of dying with pancreatic cancer than being a male. The ranking of the significant risk factors from Table 2, based on the HR, shows that the interaction between **cancer stage (Regional)** and **patient having Emphysema** ($X_{2R}X_{8Y}$) is the highest prognostic factor to the survival of pancreatic cancer, followed by patients having diabetes (X_{6Y}), and Relatives who have pancreatic cancer (X_{5Y}) is the least prognostic factor. We also provide the 95% confidence interval of the hazard ratios (HR) corresponding to the risk factors; that is,

$$P[UCL \leq HR \leq LCL] \geq 95\%$$

where UCL and LCL are the upper and lower confidence limits and we are at least 95% confident that the hazard ratios will fall into the limits. The following Table 3 provides the three popular global tests of significance which our model is based on. As, the following table shows, our proposed model (11) is *highly significant* based on all the three statistical tests.

Table 3: Global Statistical Significance of the Model

Test	Test Statistics Value	df	p-value
Likelihood Ratio Test	96.6	34	$7*10^{-8}$
Wald Test	100.8	34	$2*10^{-8}$
Score (log-rank) Test	109.9	34	$6*10^{-10}$

V. Assumptions of Cox PH Model and Validation of the Proposed Model

In order to apply the CPH model, we must verify that the following three key assumptions are satisfied, prior to its implementation. Failure to satisfy these assumptions will bring about inaccurate decisions about the subject matter.

1. **Proportional hazard (PH) assumption**: The *proportional hazard* assumption of the Cox model can be validated depending on formal statistical tests. A non-statistical significance of all risk factors along with the interactions in the model with the global test is an evidence that the PH assumption is well-grounded. Another way to verify the PH assumption is by investigating the plot of scaled Schoenfeld residuals (18) (19) against the transformed time. The Schoenfeld residuals are independent of time; a non-random pattern against time is evidence of a violation of the PH assumption. We calculate the Schoenfeld residuals for each of the risk factors and all interactions.

   The data consists of times $T_1, T_2, \ldots, T_n$ which are either observed survival times or censored times with censoring indicators $\delta_1, \delta_2, \ldots, \delta_n$. $\delta_i = 1$ implies T_i is observed, and $\delta_i = 0$ implies T_i is censored. Suppose there are p fixed covariates/risk factors $Z_1, Z_2, \ldots, Z_n$ and $\mathscr{R}_i$ be the risk set at time T_i denoted as $\mathscr{R}_i = \{j : T_j \geq T_i\}$. Given the setup, the *partial likelihood*, proposed by Cox (1975) is defined by:

$$L(\beta) = \sum_{i=1}^{n} \delta_i \Big[\beta^T Z_i - log\Big[\sum_{j \in \mathscr{R}_i} exp(\beta^T Z_j)\Big]\Big] \ . \qquad (14)$$

   Let $\hat{\beta}$ be the usual estimator of β that minimizes $L(\beta)$ in (13). Also, let $t_{(i)}$ be the i^{th} ordered observed survival time and $Z_{(i)}$ and $\mathscr{R}_i$ the corresponding covariate vector and risk set. Then Schoenfeld's residuals are defined as follows:

$$\hat{r}_i = Z_{(i)} - \frac{\sum_{j \in \mathscr{R}_i} Z_j exp(\hat{\beta}^T Z_j)}{\sum_{j \in \mathscr{R}_i} exp(\hat{\beta}^T Z_j)} \ . \qquad (15)$$

   The following Figures 5 and 6 illustrate the plot of the scaled Shoenfeld residual against time for all risk factors and interaction terms used in the model (11), respectively. It shows that there is no pattern as a function of time. Thus, the residuals are randomly scattered with no systematic departures from the horizontal fitted smoothing spline deep line (that is, the residuals are independent of times).

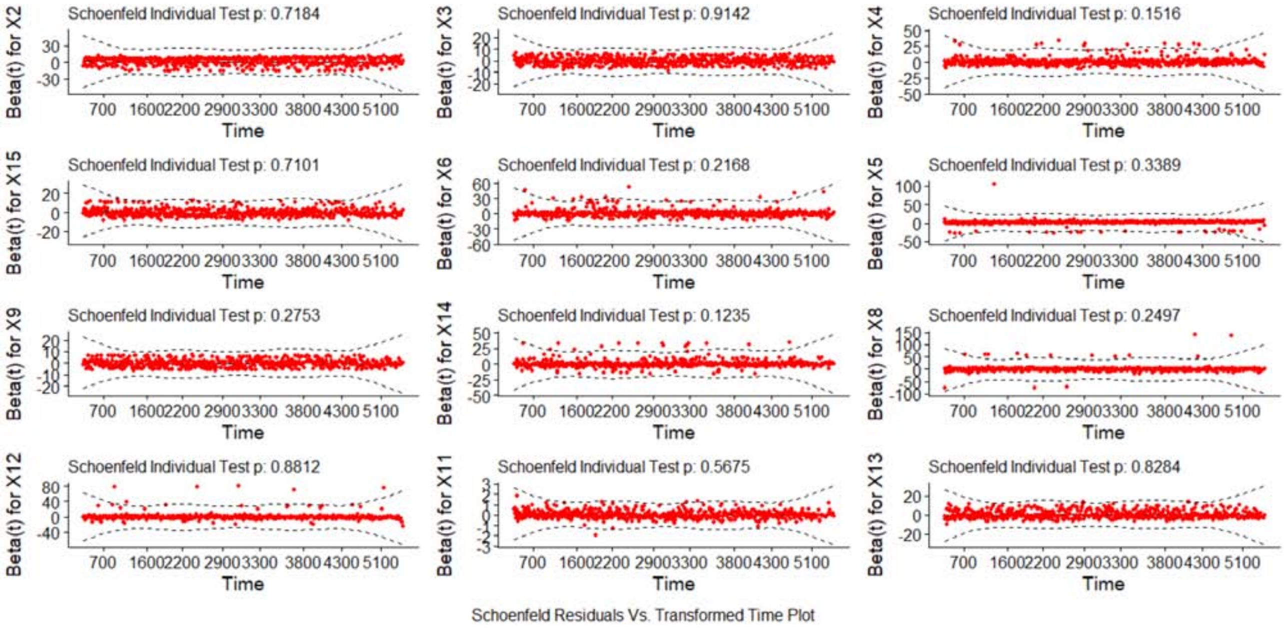

Figure 5: Testing Proportional Hazard Assumption for Individual Risk Factors

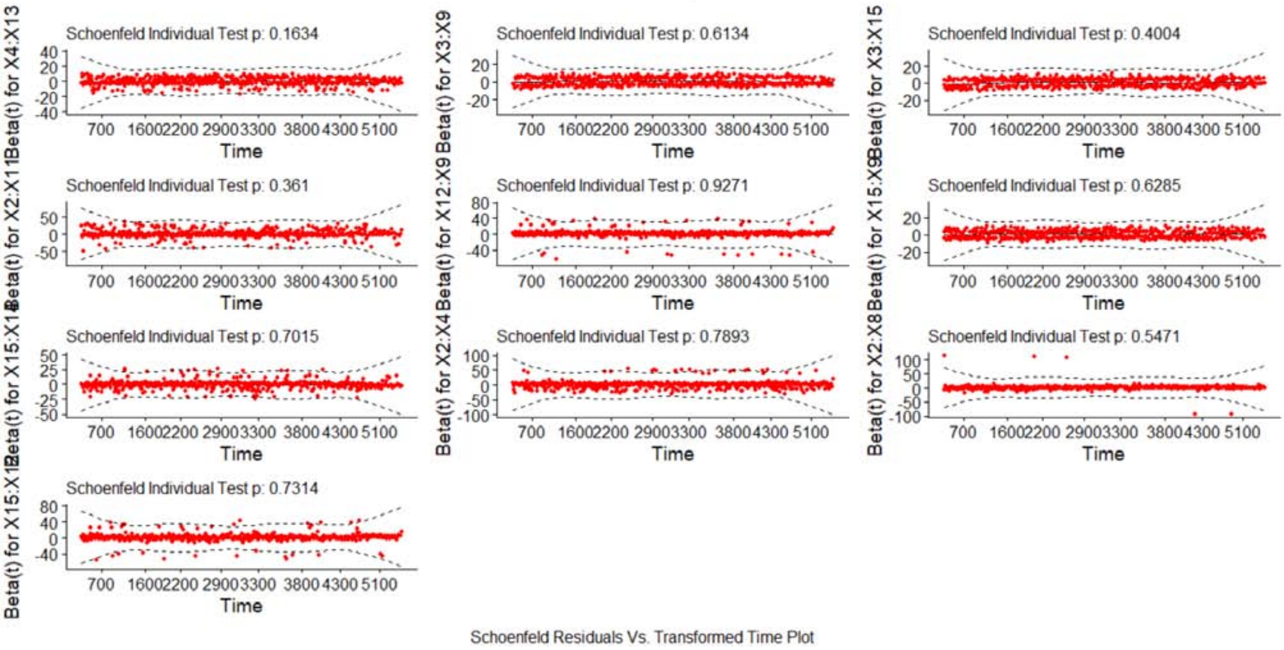

Figure 6: Testing Proportional Hazard Assumption for all Interactions

A formal test for the PH assumption is given in Table 4. The covariates and the global test are non-statistically significant given by the large p-values. This is a further justification of the validity of the PH assumption for our proposed model. We have included all fourteen risk factors and ten interaction terms in the table. The number of terms in Table 4 is greater than Table 2 since we have included all of the fourteen individual risk factors used in our analysis in Table 4.

Table 4: Testing Proportional Hazard Assumption

Risk Factors	χ^2	**p-value**
X_2	.66	.72
X_3	.01	.91
X_4	2.05	.15
X_{15}	.14	.71
X_7	3.39	.1
X_6	1.5	.21
X_8	1.3	.25
X_{12}	.02	.88
X_{14}	2.37	.12
X_{13}	.05	.82
X_{11}	.32	.56
X_{10}	2.56	.11
X_5	2.16	.34
X_9	1.19	.27
$X_4 \bigcap X_{13}$	1.94	.16
$X_3 \bigcap X_9$	.25	.61
$X_3 \bigcap X_{15}$	.71	.4
$X_2 \bigcap X_{11}$	.04	.36
$X_{12} \bigcap X_9$	.008	.93
$X_{15} \bigcap X_9$	.23	.63
$X_{15} \bigcap X_{14}$	.14	.7
$X_2 \bigcap X_4$	.47	.79
$X_2 \bigcap X_8$	1.2	.55
$X_{15} \bigcap X_{12}$	.12	.73
GLOBAL	44.17	.1

2. **Linear Functional Form of continuous Risk Factors**: Often, many researchers assume that the continuous risk factors in the Cox PH model have a linear form. However, one should verify this assumption before implementation of the model. Representing the Martingale residuals against continuous covariates is a graphical form, is a common approach to identify the nonlinearity or, in other words, to assess the functional form of a covariate. For a given continuous covariate, the plot patterns may suggest that the variable is not properly fit. Nonlinearity is not a problem for categorical risk factors. So we only investigate plots of martingale residuals against the only continuous covariate X_{11}. Sometimes, these plots can help select the appropriate functional forms of the risk factors in the Cox model. The *martingle residual*, proposed by Therneau and Grambsch (21) is given by,

$$\hat{M}_i = \delta_i - \hat{\Gamma}_0(t_i) exp\Big[\sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \sum_{j \neq k} \hat{\eta_{jk}} X_{ij} X_{ik}\Big], \;\; j \neq k. \;,$$

where δ_i denotes the event indicator for i^{th} observation, $\hat{\Gamma}_0(t_i)$ is the estimated cumulative hazard at the final follow-up time for the i^{th} observation. Martingale residuals, M_i, have a skewed distribution. We have, $\hat{M}_i = 1$ for for maximum possible values and $\hat{M}_i = -\infty$

for minimum possible values. Positive values of $\hat{M}_i$ indicate those patients expired too early compared to expected survival times. On the contrary, negative values of $\hat{M}_i$ correspond to patients who were alive for a long time.

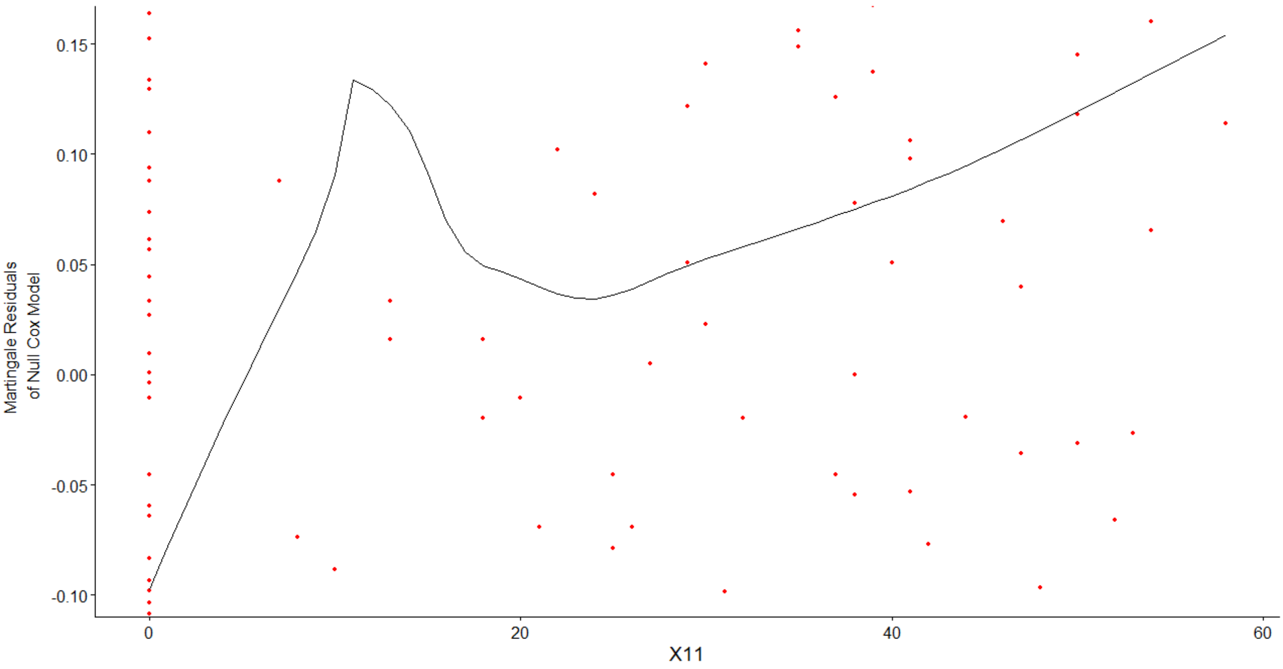

Figure 7: Validating the Linearity Assumption of the Continuous Covariate

As shown in the Figure 7, the data points are fairly linear for almost all points except around $X_{11} = 10$. The continuous covariate X_{11} is the *number of cigarette smoking years* of an individual patient. There are several patients who did not smoke at all (indicated by the points around zero). If we omit these observations, the pattern of the graph is fairly linear and increasing.

3. **Testing influential observations and Outliers**: Often influential observations can cause problems with modeling results. In order to check the influential observations, we visualized the dfbeta values. The dfbeta values estimates the influence of the i^{th} - patient observation on the regression coefficients β_j. A high value of dfbeta must be investigated carefully. Another method for checking influential observations is by assessing the *deviance residuals* (symmetric/normalized transformation of the Martingale residuals) plot. The deviance residual is defined by

$$d_i = sin(\hat{M}_i)\sqrt{2}\sqrt{-\hat{M}_i - \delta_i log(\delta_i - \hat{M}_i)}.$$

In the above equation, $\hat{M}_i$ implies $d_i = 0$. The square root shrinks the large negative martingale residuals, while the logarithm transformation expands those residuals that are close to zero. The distribution of the residuals must approximately be symmetrical around mean zero and standard deviation of one. A very large/small/distant deviance residual values indicate influential observations or outliers. Figure 8 below implies that none of the observations is exceedingly influential individually, on average.

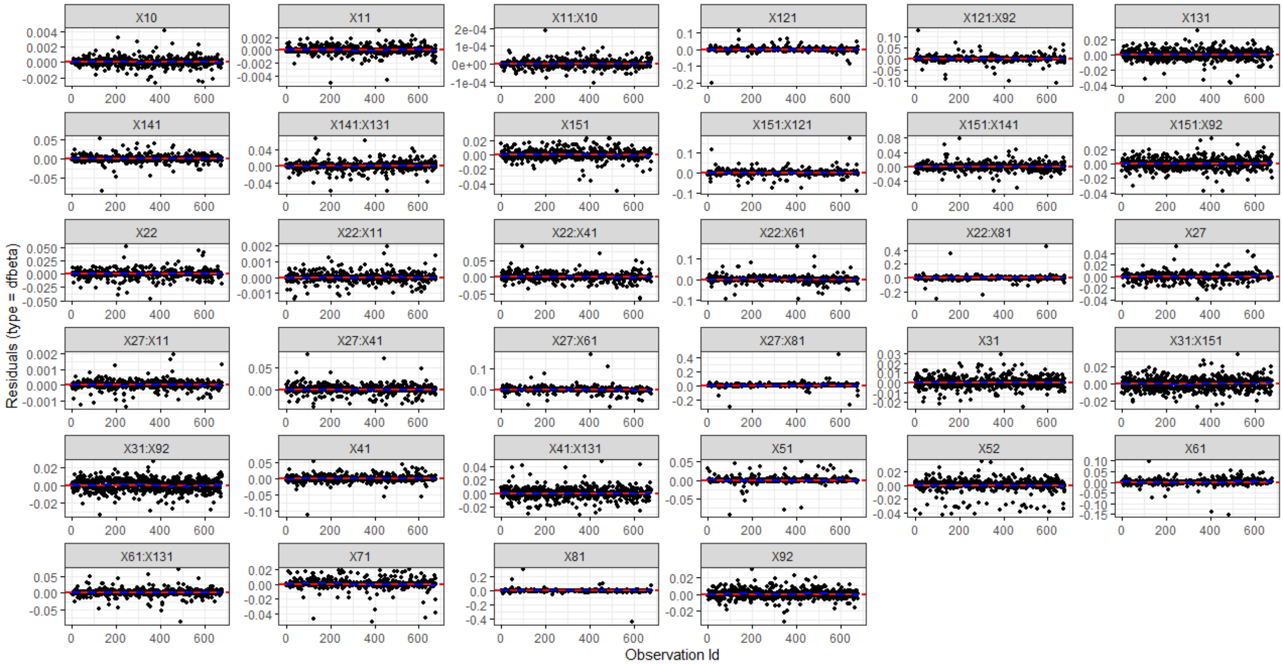

Figure 8: Assessing Influential Observations in the Model by dfbeta

The following Figure 9 plots the deviance residual and the residual pattern looks fairly symmetrical around zero. The mean deviance residual for our model is .2 which is very small.

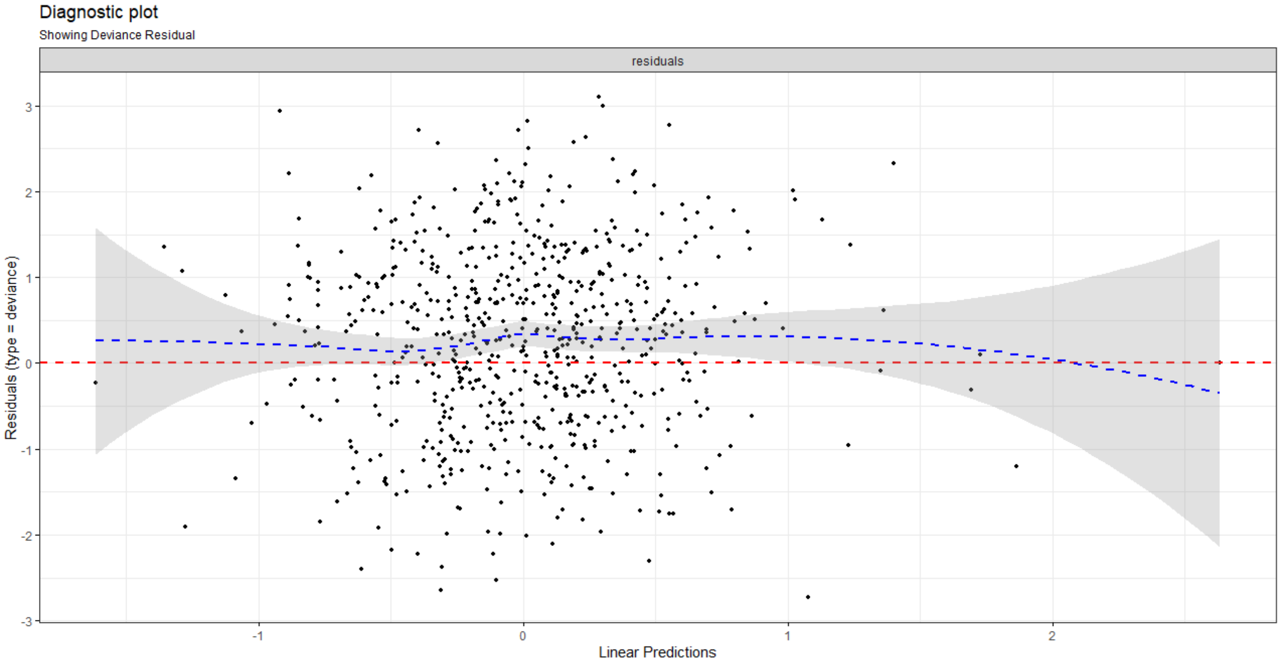

Figure 9: Assessing Influential Observations in the Model by Deviance Residual

VI. Results and Discussions

Given the risk posed by pancreatic cancer in the past few years, it is imperative to investigate the clinical diagnosis and enhance the therapeutic/treatment strategy of pancreatic cancer. The primary treatment for most types of pancreatic cancer is chemotherapy. Sometimes, with chemotherapy, specific therapy drugs are used. Usually, surgery and radiation therapy do not fall under crucial treatments for pancreatic cancer, but they might be used in exceptional circumstances. Also, the treatment approach for children with pancreatic cancer can be slightly different from that used for adults. Several research approaches and statistical methodologies (23) (24) have been developed to cure pancreatic cancer patients and boost their survival times. Chakraborty & Tsokos (to be published) performed data-driven research on pancreatic cancer patients by performing parametric analysis to improve the survival probabilities of patients of different cancer stages. In the present study, we initially investigated if there exists any statistically significant difference between the *true* mean survival times of the male and female pancreatic cancer patients using the Wilcoxon two-sample rank-sum test. The p-value (.47 > .05) of the test result suggests that there is no evidence of a significant difference between the true mean survival times of the males and the females. Hence, we proceed to perform to develop the Cox-PH (CPH) model with the combined information of male and female patients. While developing the CPH model, it is very important to justify the model assumptions. In the preliminary analysis, we found that all of the risk factors except age (X_1) did not satisfy the proportional hazard assumption. Thus, we introduced stratification in our model by dividing the covariate age into two groups. By doing stratification, we obtained more valid estimates of the other covariates, and the proportional hazard assumption was satisfied for all risk factors, including age. Performing stratification, we restrict the effect of the covariates to be the same for each stratum. Our final developed Cox-PH model given by equation (11) identified all the significant risk factors along with all the significant interaction terms as contributing to the hazard. After building our model, we proceed to rank all significant individual risk factors and all possible significant interactions according to the hazard ratio, as shown in Table 2. From Table 2, we observe that X_{6Y} (patients having diabetes), X_{4Y} (patients taking ibuprofen regularly), X_{2D} (patients who are in stage **distant** (Cancer has spread to distant parts of the body)), X_{9F} (sex), and X_{15Y} (hypertension) are the most contributing risk factors individually to the survival of patients with a hazard ratio (HR) of 2.39, 1.83, 1.63, .7, and .7, respectively. For the risk factor X_{6Y}, HR = 2.39 indicates a strong association between the patients having diabetes and increased risk of death due to pancreatic cancer. Keeping the other covariates constant, being a diabetic patient has a 2.39-fold increase in the hazard of death; that is, 2.39-fold increased risk (or decreased survival). It is important to note that according to the American Cancer Society, one of the main risk factors of pancreatic cancer is diabetes which is supported by our study. Also, we have found that those who take ibuprofen regularly have an increased risk of 1.83-fold than those who do not take the medication on a regular basis. Also, being a female has approximately 30% less hazard than a male patient. Among the most significant interactions we have $X_{2R}X_{8Y}$, $X_{15Y}X_{12Y}$, $X_{12Y}X_{9F}$, $X_{3Y}X_{15Y}$, $X_{3Y}X_{9F}$, $X_{15Y}X_{9F}$, and $X_{2R}X_{11}$ with hazard ratio 8.84, 2.28, 1.98, 1.83, 1.5, 1.37, and 1.01 respectively. The most contributing risk factor is an interaction term ($X_{2R}X_{8Y}$) (patients with emphysema and cancer stage regional with HR = 8.84). However, they do not contribute significantly to survival. We see that X_{15Y} (hypertension) has a lower risk of survival (HR = .79). However, interacting with X_{12Y} (diverticulosis), it has a hazard ratio of 2.28. Also, interacting with X_{3Y} (person who uses Aspirin Regularly), it has a hazard ratio of 2.28. It is also important to note that X_{3Y} individually has lower risk (better survival) with HR = .6. Although X_{12Y} (diverticulosis) and X_{9F} (female) has a hazard ratio less than one, their combined effect remains significant with HR = 1.98.

VII. Conclusion

In this study, we have estimated the survival probabilities of patients diagnosed with pancreatic cancer using the semi-parametric Cox proportional hazard (CPH) model. We believe the proposed Cox-PH model given by equation (11) gives an accurate estimate of the survival probability of patients diagnosed with pancreatic cancer. The stratification of the age produced more reliable estimates of the risk factor included in the CPH model. We identified seven significant risk factors and ten significant interaction terms as contributing to the survival probability of patients diagnosed with pancreatic cancer, as described in Table 2. We also ranked those risk factors and their interactions based on the hazard ratio. There have not enough studies been done in the literature that incorporates the **significant interaction effect** of two risk factors. Interaction effects play a major role as a prognostic factor in addition to the individual risk factors in the CPH model. We found some of the risk factors used in our study individually have hazard less than one, but by combining with some other risk factor, the hazard was more than 1.5, and the combined effect was significant. Our final proposed Cox-PH model is of very high quality, robust, and efficient, given by the fact that it satisfies all the major assumptions described in Section 5. The stepwise model selection procedure was utilized to carefully assess and select the risk factors and the interaction term based on their statistical significance to the survival probability. Depending on the survival analysis of the survival times based on the CPH model of the pancreatic cancer patients, we recommend the following.

1. Besides the survival time of patients, if any additional details regarding some of the potential risk factors are known, then use of the Cox proportional hazard (CPH) model can reflect a better picture of covariate effect on survival via hazard ratio.

2. Before implementing the developed CPH model, one should be careful about the fact that the CPH model assumptions are satisfied. In our present analysis, we justified the key assumptions of the CPH model.

3. The significant two-way interaction effects of the risk factors in the CPH model should not be excluded because they can significantly influence the prediction accuracy of the model and survival rate of pancreatic cancer patients, which might lead to serious clinical and therapeutic/treatment issues.

4. The ranking of the individual and interacting risk factors can be wisely used in pancreatic cancer research to improve the treatment options.

Acknowledgement

The authors are thankful to the National Cancer Institute (NIH) for making the data available.

Funding
Not Applicable

Ethics approval and consent to participate
Not Applicable

Competing interests
There are no competing interests to declare.

References Références Referencias

1. Li, D., Xie, K., Wolff, R., & Abbruzzese, J. L. (2004). Pancreatic cancer. The Lancet, 363(9414), 1049–1057. doi:10.1016/s0140-6736(04)15841-8
2. Vincent, A., Herman, J., Schulick, R., Hruban, R. H., & Goggins, M. (2011). Pancreatic cancer. The Lancet, 378(9791), 607–620. doi:10.1016/s0140-6736(10)62307-0
3. Kleeff, J., Korc, M., Apte, M. et al. Pancreatic cancer. Nat Rev Dis Primers 2, 16022 (2016). https://doi.org/10.1038/nrdp.2016.22
4. Ferlay, J. et al. GLOBOCAN 2012: cancer incidence and mortality worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer [online], http://globocan.iarc.fr (2013).
5. Cox, David R (1972). "Regression Models and Life-Tables". Journal of the Royal Statistical Society, Series B. 34 (2): 187–220. JSTOR 2985181.
6. Perera, M. and Tsokos, C. (2018) A Statistical Model with Non-Linear Effects and Non-Proportional Hazards for Breast Cancer Survival Analysis. Advances in Breast Cancer Research, 7, 65-89.

7. Asano J, Hirakawa A, Hamada C. Assessing the prediction accuracy of cure in the Cox proportional hazards cure model: an application to breast cancer data. Pharm Stat. 2014 Nov-Dec; 13(6): 357-63. doi: 10.1002/pst.1630. Epub 2014 Jul 16. PMID: 25044997.
8. Yong, X., & Tsokos, C. (2009). PROBABILISTIC COMPARISON OF SURVIVAL ANALYSIS MODELS USING SIMULATION AND CANCER DATA.
9. Du X, Li M, Zhu P, Wang J, Hou L, Li J, et al. (2018) Comparison of the flexible parametric survival model and Cox model in estimating Markov transition probabilities using real-world data. PLoS ONE 13(8): e0200807. https://doi.org/10.1371/journal.pone.0200807.
10. Kleinbaum D.G., Klein M. (2012) The Cox Proportional Hazards Model and Its Characteristics. In: Survival Analysis. Statistics for Biology and Health. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-6646-9_3
11. Kalbeisch JD, Prentice RL (1980) The Statistical Analysis of Failure Time Data. New York: John Wiley Sons 1980: 1-321.
12. Brody Tom (2011) Clinical Trials: Study Design, Endpoints and Biomarkers, Drug Safety, and FDA and ICH Guidelines. Academic Press 2011: 165-168.
13. H Akaike (1974) A new look at the statistical model identication, IEEE Trans. Autom Control 19: 716-723.
14. Sashegyi, A., & Ferry, D. (2017). On the Interpretation of the Hazard Ratio and Communication of Survival Benefit. The oncologist, 22(4), 484–486. https://doi.org/10.1634/theoncologist.2016-0198.
15. L Douglas Case, Gretchen Kimmick, Electra D Paskett, Kurt Lohman, Robert Tucker (2002) Interpreting Measures of Treatment Effect in Cancer Clinical Trials. The Oncologist 7:181-187.
16. "Theory of Partial Likelihood." Ann. Statist. 14 (1) 88 - 123, March, 1986. https://doi.org/10.1214/aos/1176349844
17. Fernandez, L. (1986). Non-parametric maximum likelihood estimation of censored regression models. Journal of Econometrics, 32(1), 35–57. doi:10.1016/0304-4076(86)90011-4.
18. GRAMBSCH, P. M., & THERNEAU, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3), 515–526. doi:10.1093/biomet/81.3.515.
19. Winnett, A. (2001). Miscellanea. A note on scaled Schoenfeld residuals for the proportional hazards model. Biometrika, 88(2), 565–571. doi:10.1093/biomet/88.2.565.
20. Mamudu L , Tsokos CP and Oluwaseun O E (2020) .Survival Analysis of Multiple Myeloma Cancer (MMC) Using the Cox-Proportional Hazard Model. Med Clin Res 5 (7):147.
21. Therneau, T., Grambsch, P., & Fleming, T. (1990). Martingale-Based Residuals for Survival Models. Biometrika, 77(1), 147-160. doi:10.2307/2336057.
22. Michaud DS. Epidemiology of pancreatic cancer. Minerva Chir. 2004 Apr;59(2) 99-111. PMID: 15238885.
23. Chakraborty, A. and Tsokos, C.P. (2021) Parametric and Non-Parametric Survival Analysis of Patients with Acute Myeloid Leukemia (AML). Open Journal of Applied Sciences, 11, 126-148. https://doi.org/10.4236/ojapps.2021.111009.
24. Aditya Chakraborty & Chris P. Tsokos. A Real Data-Driven Analytical Model to Predict Happiness. Sch J Phys Math Stat, 2021 Mar 8(3): 45-61.