Earthline Journal of Mathematical Sciences
E-ISSN: 2581-8147
Volume 14, Number 4, 2024, Pages 817-839
https://doi.org/10.34198/ejms.14424.817839

Efficiency of a New Parametric Cox Proportional Hazard Model Using Monte Carlo Simulation Study

Precious O. Ibeakuzie[1,*] and Sidney I. Onyeagu[2]

[1] Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria
e-mail: op.ibeakuzie@unizik.edu.ng

[2] Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria

Abstract

In clinical studies, statistical models have proved useful in inferential analysis. This study examines the parametric Cox PH model using simulation studies. Through simulation studies, the study demonstrated the usefulness of the parametric Cox model by comparing its statistics, such as the concordance, confidence interval, redundancy or otherwise of the covariates, and median survival time, with those of classical Cox and logistic models. By extension, the ROC was plotted to show similarity with the parametric Cox model. The results show that the parametric Cox PH model has a higher concordance ratio of 0.9810 while the classical Cox PH model has 0.7810 concordance ratio. In both model scenarios, the variable Disease-Free survival indicator did not produce any value. The mean square error of the parametric Cox PH is lower than that of the classical Cox PH model. More covariates are significant in the parametric Cox PH model than in the classical Cox PH model. This tells that the proposed parametric Cox PH model improves the classical Cox PH model. The confidence interval for both models is seemingly the same. Because the assumptions of the Cox PH model were not violated in this study, given that the exponential distribution has a constant hazard

Received: May 16, 2024; Accepted: June 13, 2024; Published: June 19, 2024
2020 Mathematics Subject Classification: 62N02.
Keywords and phrases: exponential hazard function, Cox PH model, parametric model, estimation, simulation studies.
*Corresponding author

rate, it is therefore recommended that other choices of non-constant hazard rate functions be made and deployed in the classical Cox PH model to attain some variant parametric Cox PH models.

1 Introduction

Cox Proportional Hazard also known as Cox PH model for survival analysis has come a long way in the literature with some studies both in application and software implementation being remarkable such as Allison [1,2], Cox [3,4], Fox [5], Fox and Weisberg [6], Hosmer et al. [7], Therneau [8] Therneau et al. [9]. The survival time of a particular event is called the time-to-event, Samawi et al. [10]. The time of death and time to develop a disease are examples of survival data. Statistical methods for survival analysis have been applied to many vital fields of research. Generally, survival analysis uses data to predict survival probability and identify risk and/or prognostic factors related to subjects' survival and disease progression. An essential aspect of survival data is not usually fully observed in all subjects under study, leading to different censored data types. Subjects in a study are usually assumed to be selected randomly (interred the study randomly) in the sense of simple random sample (SRS) Scheaffer et al. [11].

The Cox PH model is the most commonly used survival data analysis technique that simultaneously allows one to include and to assess the effect of multiple covariates Bradburn et al. [12]. These model covariates can include the variables of specific research interest (treatment groups), as well as potential confounders for which the researcher wants to control (demographic and other clinical factors). Multiple strategies for covariate selection have been described, and the aim of the study-most often to determine the effect of a covariate while controlling for confounding versus prediction of survival using a set of predictor variables-should be considered in choosing a strategy Clark et al. [7], Hosmer et al. [13] and Bradburn et al. [14].

The Cox PH model assumes that the hazard function for any individual at any

time is the product of an underlying baseline hazard function and an exponential function of the predictor variables. Mathematically, it can be represented as:

$$h(t|X) = h_0(t) \times \exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right), \tag{1}$$

where $h(t|X)$ is the hazard function at time t given the predictor variables X. $h_0(t)$ is the baseline hazard function representing the hazard when all predictor variables are zero. $\beta_1, \beta_2, \cdots, \beta_p$ are the regression coefficients associated with the predictor variables.

Ibeakuzie and Onyeagu [15] deployed the hazard rate function of the classical Cox PH model in designing the parametric Cox PH model. They demonstrated the usage of the new parametric Cox PH model in modeling Bone Marrow data.

Theorem 1 (**Parametric Cox PH model**). *Let X_i be predictor variables (covariates) for survival data with coefficients β_i, $i = 1, 2, \cdots, p$. Define the baseline hazard function $h_0(t)$ of the Cox PH model as the hazard function of the exponential distribution that is $h_0(t) = \lambda$, an improved Cox PH model can be constructed as*

$$h_I(t|X) = \lambda \times \exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right), \tag{2}$$

where $h_0(t) = \lambda$ is the baseline hazard function provided not all the predictors are zeros.

Essentially, Bone marrow is the spongy tissue inside some bones, including the hip and thigh bones. It contains immature cells called stem cells, which can become red or white blood cells or platelets. Many people with blood cancers, such as leukaemia and lymphoma, sickle cell anaemia, and other life-threatening conditions, rely on bone marrow or cord blood transplants to survive. The two types of bone marrow are red bone marrow, known as myeloid tissue, and yellow bone marrow, known as fatty tissue. Both types of bone marrow are enriched with blood vessels and capillaries. Bone marrow makes more than 220 billion new blood cells every day. Most blood cells in the body develop from cells in the bone marrow.

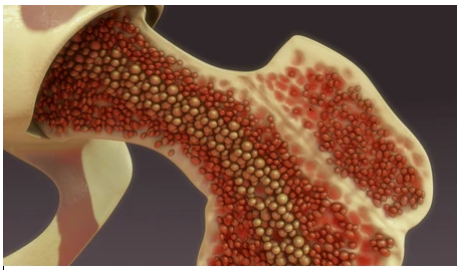

Figure 1: Human Bone Marrow.

People need healthy bone marrow and blood cells to live. When a condition or disease affects bone marrow so it can no longer function effectively, a marrow or cord blood transplant could be the best treatment option. For some people, it may be the only option. In many clinical studies, subjects are followed up repeatedly, and response data is collected, for example, a biomarker. The time of an event is also usually of interest, which might be explicit, e.g., death, or implicit, e.g., dropout. Bone marrow transplantation (BMT) is the treatment of choice for many leukaemias, solid tumours, and metabolic diseases. The field of bone marrow research is highly dependent on in vivo experimentation because in vitro techniques do not mimic these complicated in vivo systems, Duran-Struuck and Dysko [16]. Historically, the field of bone marrow transplantation (BMT) has been highly dependent on in vivo models. Regarding numbers, the mouse is the mammal used most frequently for BMT studies. Murine models have clear advantages in their physiologic and pathologic traits similarities with other mammals, including humans. The small mass of mice, their large litter sizes, short pregnancy period, availability of diverse stocks and strains, and transgenic, knockout, and knock-in lines have made them one of the most valuable and versatile experimental animal models for human and veterinary biomedical research. During BMT, recipient mice may receive a genetically identical bone marrow graft or, often, a genetically disparate graft. If genetically disparate BM grafts are transplanted, a severe immune reaction from the donor cells attacks the hosts' tissues. However, if the host immune system is not

pretreated (that is, immunosuppressed to some degree), failure of engraftment or graft rejection (of the donor BM) may occur. Many methods are used to ablate the immune system. The easiest and most commonly used method experimentally is total-body irradiation (TBI), which is achieved by placing the mice in specifically designed irradiators; the dose of whole-body gamma irradiation causes the animals to become either transiently or chronically immunosuppressed. Because of the animal's weakened immune system, strict veterinary and husbandry care requirements are needed to ensure the well-being of these animals.

More precisely, the process of BMT should be termed hematopoietic or hematopoietic stem cell transplantation because the stem cells responsible for reconstituting the immune system can now be harvested directly from the circulation. Currently, most transplants deliver peripheral-blood-mobilized stem cells, not cells harvested directly from the BM by aspiration. Another source of stem cells used currently is the umbilical cord, Urbano-Ispizua [17]. During BMT, a donor inoculum is given to a recipient. The inoculum contains pluripotent hematopoietic stem cells and more mature hematopoietic cells arising from the myeloid, lymphoid, and erythroid lineages. These hematopoietic cells are harvested from bone marrow (for example, the iliac crest or long bones) or from the circulation after administration of granulocyte colony-stimulating factor or other growth factors that mobilize these cells to the peripheral circulation. In small animal models such as mice, the bone marrow from a donor mouse is the most common source of stem cells; however, in larger species such as dogs, pigs, and primates, peripheral blood stem cells can be harvested more easily due to the greater blood volume of the animals. Two types of progenitors reconstitute the recipient's immune system after hematopoietic stem cell transplantation: short-term and long-term hematopoietic cell progenitors.

Cancer has become an impediment to human longevity and high quality of life. According to the latest global cancer statistics from GLOBOCAN 2020, colorectal cancer ranked third in terms of new cases, with 1.93 million accounting for 10% of all new cancers, and second in terms of deaths, with 940,000 accounting for 9.4% of all deaths due to cancer in that year, Sung et al. [18]. This time-to-event outcome

may censor the longitudinal data. Due to the importance of this medical condition, many researchers have investigated Cancer vis-á-vis bone marrow from different perspectives. For instance, Transverse colon cancer has been reported to account for about 10% of colorectal cancers, West et al. [19]. The transverse colon is in a special high position between the ascending and descending colon, in the middle and anterior part of the entire colon, excluding the hepatic and splenic flexures. It has a maximum length of about 50 cm. Since the transverse colon differs from the rest of the colon in terms of embryonic development, anatomical structure, blood supply, and pathogenetic characteristics, it is necessary to delineate the different segments of the colon clearly and to provide precise and individualized treatments according to the specific characteristics of the transverse colon, which is also in line with contemporary medical concepts. However, most studies on colorectal cancer have focused on the ascending and descending colon, which have obvious differences. The transverse colon, the link between the two, has received little attention in research. Adenocarcinoma arises from the glandular epithelium, ducts, or secretory epithelium and is characterized by adenoid structure formation. It is the most common clinical type of colon cancer, accounting for 90-95% of cases, and has a better prognosis than other pathological types. Research on adenocarcinoma of the transverse colon (ATC) would help improve the clinical outcomes, Su et al. [20].

Modelling the longitudinal (data that is collected through a series of repeated observations of the same subjects over some extended time frame) and event-time outcomes separately, for example, using linear mixed models Laird and Ware [21] or Cox regression models Cox [3] can therefore be inefficient, and can lead to biased effect size estimates if the two outcome processes are correlated Ibrahim et al. [22]. Research into joint modelling of longitudinal and time-to-event data has received considerable attention during the past two decades Ibrahim et al. [22], Asar et al. [23], Henderson et al. [24], Rizopoulos [25] and Wulfsohn and Tsiatis [26].

The motivation for this study are

1. To estimate the parametric Cox PH structure parameters using the classical

maximum likelihood estimation.

2. To compare the parametric Cox PH structure with the classical Cox PH and logistic regression in simulation studies.

3. To Compare the parametric Cox PH structure with the classical Cox PH regression model using simulation studies.

2 Simulation Studies

Parametric simulation offers numerous benefits over traditional cost estimation methods. Here are some key advantages: - Improved accuracy: Parametric simulation considers a wide range of variables and their interactions, resulting in more realistic and accurate cost estimates. A Monte Carlo simulation is adopted for sample sizes $n = 25, 50, 100, 150, 200$ at 1000 replication. The parameters are estimated using the maximum partial likelihood iterative algorithm in R software for the parametric Cox PH model.

- Define a score vector

$$U(\beta) = \left(\frac{\partial \ell_p(\beta)}{\partial \beta_1}, \cdots, \frac{\partial \ell_p(\beta)}{\partial \beta_k} \right)'$$

  and the $k \times k$ information matrix $I(\beta)$ whose $(I, j)th$ element is $\frac{\partial^2 \ell_p(\beta)}{\partial \beta_i \partial \beta_j}$.

- Get the first approximation $\beta^{(1)} = \beta^{(0)} + I^{-1}\left(\beta^{(0)}\right) U\left(\beta^{(0)}\right)$. Second approximation $\beta^{(2)} = \beta^{(1)} + I^{-1}\left(\beta^{(1)}\right) U\left(\beta^{(1)}\right)$ and so on.

- The iterative method will converge at $(r+1)th$. Stop if $\beta^{(r)}$ and $\beta^{(r+1)}$ agree upto certain decimal places and then ML estimates $\hat{\beta} = \beta^{(r)}$ or $\beta^{(r+1)}$.

- Further $\beta^{(0)}$ from $\hat{\beta}$ more is r i.e. less likely is the convergence to $\hat{\beta}$.

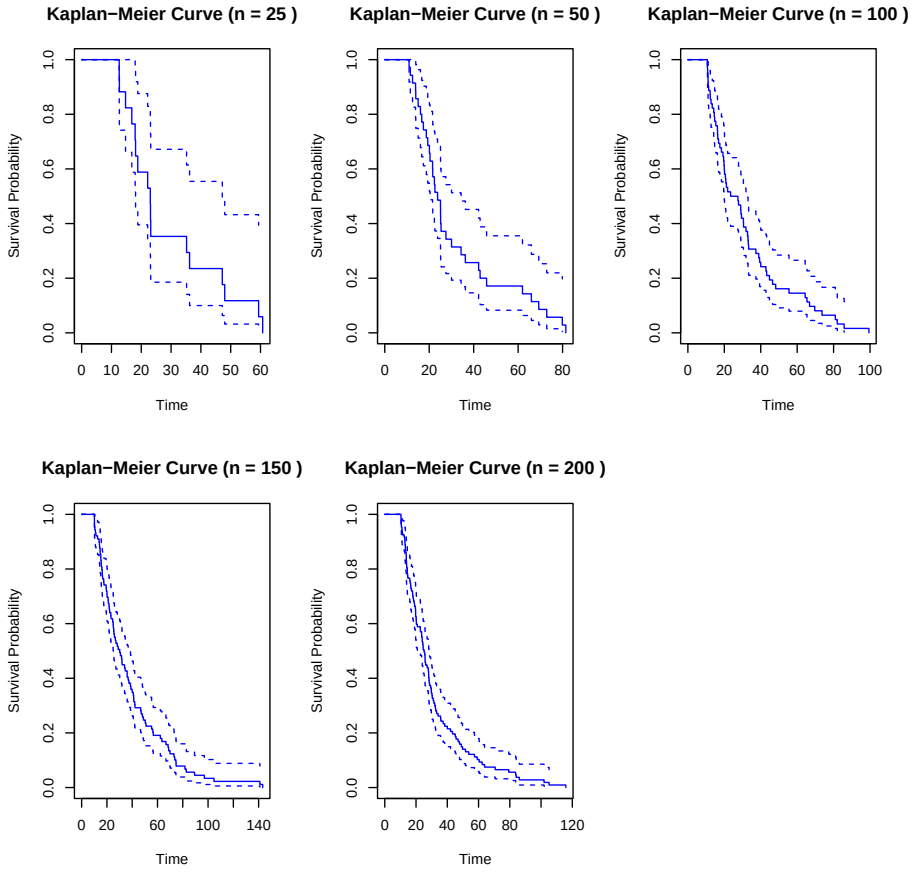

Figure 2: Kaplan-Meier curves for the simulated parametric Cox PH model.

Notice that the confidence intervals for various sample sizes in Figure 2 are wider than in Figure 3. These indicate that the parametric Cox PH model considerably provides more allowance for error in estimates than the classical Cox PH model. Though narrow estimates represent better precision, however, precision is vague if it predicates error tolerance. So, the trade-off is aptly the level of tolerance or confidence interval produced by the Kaplan-Meier curves. On this basis, the proposed parametric Cox PH model performs better than the

classical Cox PH model using the simulation result. A quick look at Figure 4 reveals a lot, and a thorough discussion will justify the comparison of the logistic regression with the proposition. ROC curves in logistic regression are used for determining the best cutoff value for predicting whether a new observation is a "failure" (0) or a "success" (1). First, let's cover what a classification cutoff is doing. When you choose a classification cutoff (let us say you choose 0.5), you are saying that you would like to classify every observation with a predicted probability from the model equal to or greater than 0.5 as a "success". Note that you will classify observations meeting this criterion as a success regardless of whether the outcome was observed to be a success. Your observed outcome in logistic regression can only be 0 or 1. The predicted probabilities from the model can take on all possible values between 0 and 1. So, for a given observation, the predicted probability from the model may have been 0.51 (51% probability of success), but your observation was a 0 (not a success).

Each dot on the curve in Figure 4 represents a different possible cutoff value for classifying predicted values. You could feasibly pick any value between 0 and 1 as the cutoff, but doing this manually for every possible meaningful cutoff value would be exhausting. So what a ROC curve does is look at every possible cutoff value that results in a change of classification of any observation in your data set (if stepping the classification cutoff up from 0.5 to 0.6 does not result in a change in how the observations are classified, well then it is not an interesting step to consider). A dot is placed on the plot for every classification cutoff that results in a classification change. Notice that there are no identifiable points on each of the AOCs. Sure, the points make up the curves since, from basic geometry, a curve is a set of collinear points.

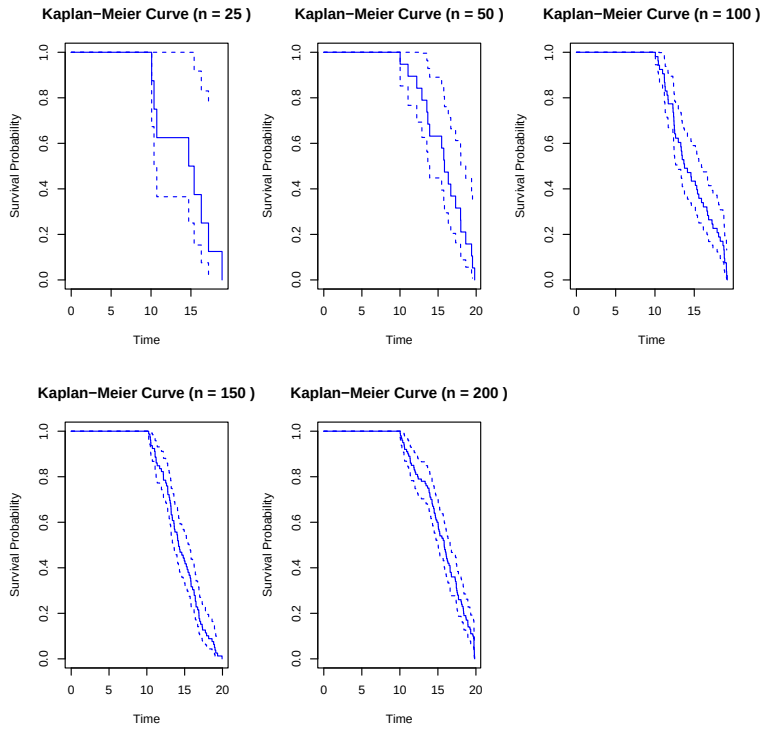

Figure 3: Kaplan-Meier curves for the simulated classical Cox PH model.

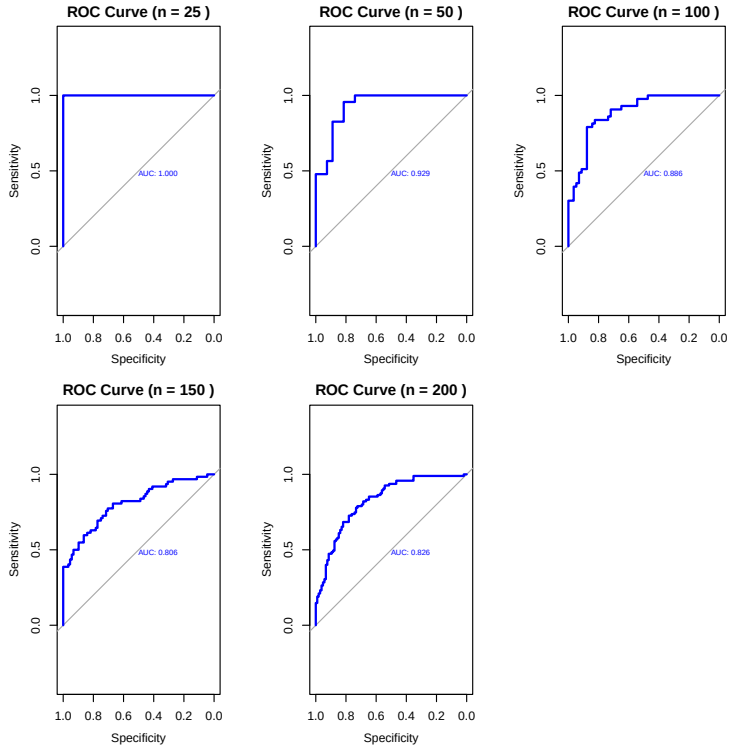

Figure 4: AOC for the simulated logistic regression model.

Whatever cutoff you choose, a certain number of the rows of data will be correctly classified (you predicted the correct value for that row), and a certain number will be misclassified. Sensitivity and specificity are two metrics for evaluating the proportion of true positives and true negatives, respectively. In other words, sensitivity is the proportion of 1s you correctly identified as 1s using that particular cutoff value or the true positive rate. Conversely, specificity is the proportion of 0s you correctly identified as 0s or the true negative rate.

ROC with the cutoff produces an area of coverage with probabilities indicated. The area can be viewed as a confidence interval of the parameter estimates. One can easily see how wide these regions are in Figure 4, even better than the proposed

parametric Cox PH model. However, due to the specificity of objectives underlying the Cox PH model vis-á-vis its parametric counterpart, this comparison cannot be extended when the real data set is deployed to demonstrate the utility of the proposed model. In statistical inference, it is important to closely study the model's behaviour before choosing the best model so that the one's choice will not mislead the users. On the above premise, a plot of the MSE of the logistic regression is made in Figure 5, and this reveals that the model is not good enough in the face of a large sample since, as the sample size increases, the mean square error (MSE) also increases.

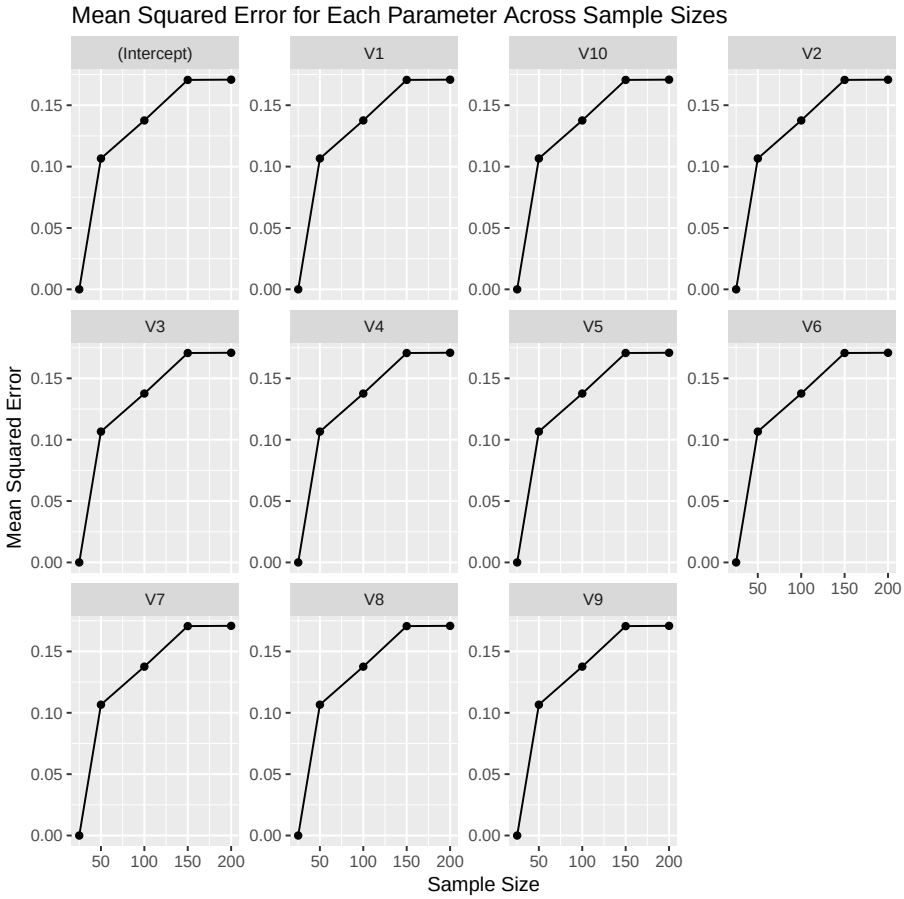

Figure 5: MSE for the simulated logistic regression model.

Table 1: Simulation result for the parametric Cox PH model.

n	covariates	β_j	e^{coefs}	e^{-coefs}	L-95	U-95	LR	Wald	p-value	con.	MSE
25	X_1	0.6786	1.9712	0.5073	0.5839	6.6544	1.6891	1.6891	0.2743	0.8603	61.5595
	X_2	-0.8654	0.4209	2.3760	0.0937	1.8907			0.2589		
	X_3	-1.0048	0.3661	2.7314	0.1002	1.3378			0.1286		
	X_4	0.2929	1.3403	0.7461	0.3632	4.9462			0.6602		
	X_5	-1.2953	0.2738	3.6522	0.0657	1.1413			0.0753		
	X_6	0.9535	2.5949	0.3854	0.6045	11.1384			0.1996		
	X_7	1.0240	2.7844	0.3592	0.5445	14.2379			0.2187		
	X_8	0.0357	1.0364	0.9649	0.2323	4.6237			0.9627		
	X_9	-0.5851	0.5571	1.7952	0.1663	1.8664			0.3429		
	X_{10}	-1.5228	0.2181	4.5849	0.0345	1.3805			0.1058		
50	X_1	-0.1766	0.8381	1.1932	0.5608	1.2524	0.9043	0.9043	0.3888	0.7765	55.0041
	X_2	-1.0451	0.3517	2.8436	0.1927	0.6416			0.0007		
	X_3	0.0892	1.0933	0.9147	0.7502	1.5933			0.6426		
	X_4	-0.5060	0.6029	1.6586	0.3788	0.9597			0.0329		
	X_5	0.1971	1.2178	0.8212	0.7297	2.0325			0.4509		
	X_6	-0.4876	0.6141	1.6285	0.3543	1.0643			0.0822		
	X_7	0.0769	1.0799	0.9260	0.7278	1.6025			0.7025		
	X_8	-0.0215	0.9787	1.0217	0.6179	1.5504			0.9271		
	X_9	0.1114	1.1178	0.8946	0.6487	1.9263			0.6883		
	X_{10}	0.6665	1.9475	0.5135	1.3087	2.8982			0.0010		
100	X_1	0.2180	1.2436	0.8041	0.9324	1.6588	2.3906	2.3906	0.1379	0.6029	47.1506
	X_2	0.1476	1.1590	0.8628	0.8277	1.6228			0.3903		
	X_3	-0.3366	0.7142	1.4001	0.5348	0.9538			0.0226		
	X_4	0.0472	1.0483	0.9539	0.8117	1.3540			0.7176		
	X_5	-0.0720	0.9306	1.0746	0.6837	1.2665			0.6472		
	X_6	-0.1059	0.8996	1.1117	0.6917	1.1698			0.4296		
	X_7	0.1597	1.1731	0.8524	0.8786	1.5664			0.2791		
	X_8	0.0466	1.0477	0.9545	0.7869	1.3950			0.7497		
	X_9	-0.0444	0.9566	1.0454	0.7117	1.2858			0.7688		
	X_{10}	-0.0198	0.9804	1.0200	0.7220	1.3313			0.8990		
150	X_1	-0.0645	0.9375	1.0667	0.7699	1.1416	0.4552	0.4552	0.5207	0.5963	42.3309
	X_2	0.0522	1.0536	0.9491	0.8516	1.3036			0.6305		
	X_3	0.0809	1.0843	0.9223	0.8433	1.3940			0.5280		
	X_4	-0.1491	0.8615	1.1608	0.6585	1.1270			0.2767		
	X_5	-0.0892	0.9147	1.0933	0.7039	1.1886			0.5046		
	X_6	-0.0284	0.9720	1.0288	0.7567	1.2486			0.8242		
	X_7	0.0609	1.0628	0.9409	0.8184	1.3801			0.6478		
	X_8	-0.1980	0.8204	1.2190	0.6593	1.0208			0.0759		
	X_9	-0.0263	0.9740	1.0267	0.7715	1.2298			0.8249		
	X_{10}	0.1100	1.1163	0.8958	0.8698	1.4327			0.3874		

From the result in Table 1, the MSE reduces as the sample size increases. It is the fundamental property of all machine learning algorithms to improve performance with more available training data. If you only have a few dozen cases to test your model, the performance will likely depend on the particular train/test split you performed. This implies that the estimation precision improves in the case of the parametric Cox PH model than that of classical Cox PH and logistic models; see also Tables 2 to 6.

Furthermore, the total number of Concordant pairs is counted and divided by the total number of pairs. This will give us the value of the concordance ratio. The higher the concordance ratio, the better the model. Comparing the parametric Cox PH with the classical Cox PH models using Tables 1 to 4, it is obvious that the proposed parametric Cox PH model is better than the classical Cox PH for large sample sizes.

Table 2: Simulation results for the parametric Cox PH model continues.

n	covariates	β_j	e^{coefs}	e^{-coefs}	L-95	U-95	LR	Wald	p-value	con.	MSE
200	X_1	-0.0151	0.9850	1.0153	0.7630	1.2715	0.3047	0.3047	0.9075	0.5943	29.4991
	X_2	0.0925	1.0969	0.9116	0.8848	1.3600			0.3989		
	X_3	-0.2105	0.8102	1.2342	0.6536	1.0043			0.0548		
	X_4	-0.0222	0.9780	1.0225	0.7893	1.2118			0.8390		
	X_5	-0.0338	0.9668	1.0344	0.7823	1.1947			0.7544		
	X_6	-0.1569	0.8548	1.1698	0.6771	1.0793			0.1872		
	X_7	-0.0822	0.9211	1.0857	0.7289	1.1640			0.4913		
	X_8	0.0875	1.0914	0.9162	0.8839	1.3477			0.4163		
	X_9	-0.0294	0.9710	1.0298	0.7623	1.2370			0.8119		
	X_{10}	0.0314	1.0319	0.9691	0.8573	1.2421			0.7396		

Table 3: Simulation results for Cox PH model.

n	covariates	β_j	e^{coefs}	e^{-coefs}	L-95	U-95	LR	Wald	p-value	con.	MSE
25	X_1	77.4377	42.0000	0.0000	0.0000	Inf	0.7012	0.7012	0.9770	1.00	127.56
	X_2	-75.5699	0.0000	660.0000	0.0000	Inf			0.9760		
	X_3	-1.3786	0.2519	3.9695	0.0000	Inf			0.9996		
	X_4	62.8667	20.0000	0.0000	0.0000	Inf			0.9708		
	X_5	-123.4589	0.0000	414.0000	0.0000	Inf			0.9538		
	X_6	77.0365	28.0000	0.0000	0.0000	Inf			0.9720		
	X_7	-57.6077	0.0000	104.0000	0.0000	Inf			0.9711		
	X_8	-202.9387	0.0000	136.0000	0.0000	Inf			0.9200		
	X_9	-60.6836	0.0000	226.0000	0.0000	Inf			0.9747		
	X_{10}	239.8216	14.0000	0.0000	0.0000	Inf			0.9314		
50	X_1	0.3786	1.4602	0.6848	0.7735	2.7566	0.9621	0.9621	0.2429	0.68	105.88
	X_2	0.0436	1.0446	0.9573	0.6414	1.7013			0.8608		
	X_3	-0.6242	0.5357	1.8667	0.3070	0.9348			0.0280		
	X_4	-0.4404	0.6438	1.5533	0.3509	1.1810			0.1549		
	X_5	0.2325	1.2617	0.7926	0.5912	2.6928			0.5478		
	X_6	0.3548	1.4260	0.7013	0.6020	3.3775			0.4199		
	X_7	0.2612	1.2985	0.7701	0.8088	2.0848			0.2795		
	X_8	-0.4552	0.6343	1.5765	0.3374	1.1924			0.1575		
	X_9	0.2511	1.2854	0.7780	0.8003	2.0645			0.2990		
	X_{10}	-0.5589	0.5718	1.7488	0.3355	0.9746			0.0399		
100	X_1	0.1294	1.1381	0.8786	0.8363	1.5489	0.0001	0.0001	0.4106	0.64	122.56
	X_2	-0.2804	0.7555	1.3237	0.5091	1.1211			0.1639		
	X_3	0.1327	1.1419	0.8757	0.7947	1.6408			0.4731		
	X_4	-0.2856	0.7516	1.3306	0.5490	1.0288			0.0746		
	X_5	-0.2306	0.7941	1.2593	0.4999	1.2614			0.3288		
	X_6	0.3381	1.4023	0.7131	0.9264	2.1226			0.1099		
	X_7	-0.1916	0.8256	1.2112	0.5754	1.1845			0.2981		
	X_8	-0.0320	0.9685	1.0325	0.6878	1.3638			0.8547		
	X_9	-0.3149	0.7299	1.3701	0.5135	1.0374			0.0792		
	X_{10}	-0.1006	0.9043	1.1058	0.6372	1.2834			0.5734		
150	X_1	-0.0706	0.9319	1.0731	0.6962	1.2474	0.5197	0.5197	0.6353	0.59	112.76
	X_2	-0.1576	0.8542	1.1707	0.6236	1.1700			0.3262		
	X_3	-0.2653	0.7670	1.3038	0.5577	1.0549			0.1028		
	X_4	0.2184	1.2441	0.8038	0.9612	1.6103			0.0970		
	X_5	-0.1481	0.8623	1.1596	0.6588	1.1287			0.2808		
	X_6	-0.0360	0.9647	1.0366	0.7086	1.3133			0.8192		
	X_7	0.1024	1.1079	0.9026	0.8811	1.3930			0.3806		
	X_8	-0.0233	0.9769	1.0236	0.7728	1.2350			0.8452		
	X_9	-0.0433	0.9576	1.0442	0.7417	1.2364			0.7399		
	X_{10}	0.0882	1.0922	0.9155	0.8203	1.4544			0.5459		

Table 4: Simulation results for Cox PH model continues.

n	covariates	β_j	e^{coefs}	e^{-coefs}	L-95	U-95	LR	Wald	p-value	con.	MSE
200	X_1	-0.0715	0.9310	1.0741	0.7437	1.1653	0.3154	0.3154	0.5324	0.58	113.36
	X_2	-0.0373	0.9634	1.0380	0.7749	1.1978			0.7370		
	X_3	-0.0398	0.9610	1.0406	0.7949	1.1619			0.6813		
	X_4	0.0474	1.0485	0.9538	0.8241	1.3340			0.7000		
	X_5	0.1379	1.1478	0.8712	0.9226	1.4280			0.2161		
	X_6	0.0255	1.0259	0.9748	0.7966	1.3211			0.8431		
	X_7	-0.0984	0.9063	1.1034	0.7285	1.1273			0.3768		
	X_8	0.1443	1.1552	0.8657	0.9148	1.4588			0.2256		
	X_9	0.0394	1.0402	0.9614	0.8616	1.2557			0.6820		
	X_{10}	-0.1380	0.8711	1.1480	0.6998	1.0842			0.2164		

Table 5: Simulation results for logistic regression.

n	covariates	β_j	MSE	p-value	con.	AUC	Odds ratios
	(Intercept)	8.68051		0.99990	1.00000	1.00000	5887.02500
	X_1	34.24979		0.99975	0.92915	0.92915	749.00000
	X_2	47.20628		0.99981	0.88576	0.88576	3.00000
	X_3	53.08038		0.99969	0.80645	0.80645	113.00000
	X_4	35.15189		0.99969	0.82627	0.82627	185.00000
25	X_5	59.22412	0.00000	0.99954	1.00000	1.00000	52.00000
	X_6	43.94169		0.99949	0.92915	0.92915	121.00000
	X_7	5.62246		0.99995	0.88576	0.88576	276.56800
	X_8	8.67823		0.99989	0.80645	0.80645	5873.63700
	X_9	3.31478		0.99998	0.82627	0.82627	27.51642
	X_{10}	56.16050		0.99954	1.00000	1.00000	246.00000
	(Intercept)	0.00236		0.99632	0.92915	0.92915	1.00237
	X_1	0.57602		0.21549	0.88576	0.88576	1.77895
	X_2	0.68382		0.25198	0.80645	0.80645	1.98143
	X_3	1.27006		0.03338	0.82627	0.82627	3.56105
	X_4	0.86462		0.09591	1.00000	1.00000	2.37410
50	X_5	0.73476	0.10662	0.24738	0.92915	0.92915	2.08499
	X_6	0.57201		0.27501	0.88576	0.88576	1.77182
	X_7	0.86884		0.18723	0.80645	0.80645	2.38413
	X_8	0.82376		0.19414	0.82627	0.82627	2.27905
	X_9	0.71369		0.17877	1.00000	1.00000	2.04151
	X_{10}	0.94881		0.06223	0.92915	0.92915	2.58263
	(Intercept)	-0.53049		0.08411	0.88576	0.88576	0.58832
	X_1	0.73491		0.02106	0.80645	0.80645	2.08529
	X_2	0.76254		0.03319	0.82627	0.82627	2.14371
	X_3	0.37477		0.19535	1.00000	1.00000	1.45466
	X_4	0.53135		0.06696	0.92915	0.92915	1.70123
100	X_5	1.21547	0.13762	0.00096	0.88576	0.88576	3.37189
	X_6	1.07398		0.00347	0.80645	0.80645	2.92701
	X_7	0.66229		0.03958	0.82627	0.82627	1.93923
	X_8	0.42158		0.12796	1.00000	1.00000	1.52436
	X_9	0.43350		0.16340	0.92915	0.92915	1.54264
	X_{10}	0.78122		0.02595	0.88576	0.88576	2.18412
	(Intercept)	-0.48238		0.01934	0.80645	0.80645	0.61731
	X_1	0.38372		0.03244	0.82627	0.82627	1.46773
	X_2	0.27390		0.17730	1.00000	1.00000	1.31509
	X_3	0.53356		0.01684	0.92915	0.92915	1.70499
	X_4	0.44024		0.03531	0.88576	0.88576	1.55308
150	X_5	0.50765	0.17064	0.01484	0.80645	0.80645	1.66138
	X_6	0.33315		0.15416	0.82627	0.82627	1.39536
	X_7	0.52399		0.02550	1.00000	1.00000	1.68876
	X_8	0.40714		0.07107	0.92915	0.92915	1.50252
	X_9	0.55804		0.01276	0.88576	0.88576	1.74724
	X_{10}	0.60001		0.00829	0.80645	0.80645	1.82213

Table 6: Simulation results for logistic regression continues.

n	covariates	β_j	MSE	p-value	con.	AUC	Odds ratios
	(Intercept)	-0.05497		0.75407	0.82627	0.82627	0.94652
	X_1	0.18612		0.28712	1.00000	1.00000	1.20457
	X_2	0.27986		0.11214	0.92915	0.92915	1.32295
	X_3	0.16563		0.36128	0.88576	0.88576	1.18013
	X_4	0.53927		0.00454	0.80645	0.80645	1.71476
200	X_5	0.19542	0.17085	0.30814	0.82627	0.82627	1.21582
	X_6	0.65330		0.00077	1.00000	1.00000	1.92187
	X_7	0.44013		0.02187	0.92915	0.92915	1.55291
	X_8	0.48515		0.01066	0.88576	0.88576	1.62442
	X_9	0.82019		0.00004	0.80645	0.80645	2.27092
	X_{10}	0.32600		0.07666	0.82627	0.82627	1.38541

3 Discussion of Results

1. We utilized the hazard rate of the exponential distribution used as the baseline hazard function in the classical Cox PH model to produce what is now known as the parametric Cox PH model given as $h_I(t|X) = \lambda \times \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$ where $h_0(t) = \lambda$ is the baseline hazard function provided not all the predictors are zeros. This model was proposed by [15].

2. From the simulation results, the confidence intervals for various sample sizes in Figure 2 are wider than in Figure 3. These indicate that the parametric Cox PH model considerably provides more allowance for error in estimates than the classical Cox PH model. Though narrow estimates represent better precision, however, precision is vague if it predicates error tolerance. So, the trade-off is aptly the level of tolerance or confidence interval produced by the Kaplan-Meier curves. On this basis, the proposed parametric Cox PH model performs better than the classical Cox PH model using the simulation result.

3. It was observed from the ROC of the logistic regression that the regions' width in Figure 4 are better than the proposed parametric Cox PH model. However, due to the specificity of objectives underlying the Cox PH model vis-á-vis its parametric counterpart, this comparison cannot be extended when the real data set is deployed to demonstrate the utility of the proposed model. In statistical inference, it is important to closely study the model's behaviour before choosing the best model so that the one's choice will not mislead the users. On the above premise, a plot of the MSE of the logistic regression is made in Figure 5, and this reveals that the model is not good enough in the face of a large sample since, as the sample size increases, the mean square error (MSE) also increases.

4. From the result in Table 1, the MSE reduces as the sample size increases. It is the fundamental property of all machine learning algorithms to improve performance with more available training data. If you only have a few dozen cases to test your model on, the performance will likely depend on the particular train/test split you performed. This implies that the estimation precision improves in the case of the parametric Cox PH model than that of classical Cox PH and logistic models; see also Tables 2 to 6.

5. Furthermore, the total number of Concordant pairs is counted and divided by the total number of pairs. This will give us the value of the concordance ratio. The higher the concordance ratio, the better the model. Comparing the parametric Cox PH with the classical Cox PH models using Tables 1 to 4, it is obvious that the proposed parametric Cox PH model is better than the classical Cox PH for large sample sizes.

6. Because the assumptions of the Cox PH model were not violated in this study, given that the exponential distribution has a constant hazard rate, it is therefore recommended that other choices of non-constant hazard rate functions be made and deployed in the classical Cox PH model to attain some variant parametric Cox PH models.

References

[1] Allison, P. D. (2010). *Survival analysis using SAS: a practical guide*. Sas Institute.

[2] Allison, P. D. (2018). Event history and survival analysis. In *The reviewer's guide to quantitative methods in the social sciences* (pp. 86-97). Routledge. https://doi.org/10.4324/9781315755649-7

[3] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[4] Cox, D. R. (2018). *Analysis of survival data*. Chapman and Hall/CRC.

[5] Fox, J. (2016). *Using the R commander: a point-and-click interface for R*. Chapman and Hall/CRC.

[6] Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage Publications. https://doi.org/10.32614/CRAN.package.carData

[7] Hosmer Jr, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: regression modeling of time-to-event data* (Vol. 618). John Wiley & Sons. https://doi.org/10.1002/9780470258019

[8] Therneau, T. M. (1997). Extending the Cox model. In *Proceedings of the first Seattle symposium in biostatistics: survival analysis* (pp. 51-84). Springer. https://doi.org/10.1007/978-1-4684-6316-3_5

[9] Therneau, T. et al. (2015). A package for survival analysis in S. In R package version 2.7, pages 2014.

[10] Samawi, H., Yu, L., & Yin, J. (2023). On Cox proportional hazards model performance under different sampling schemes. *Plos One, 18*(4), e0278700. https://doi.org/10.1371/journal.pone.0278700

[11] Scheaffer, R. L., Mendenhall, W., Ott, L., & Gerow, K. (1990). *Elementary survey sampling* (Vol. 501). Duxbury Press California.

[12] Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. *British Journal of Cancer, 89*(3), 431-436. https://doi.org/10.1038/sj.bjc.6601119

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

[14] Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part III: multivariate data analysis–choosing a model and assessing its adequacy and fit. *British Journal of Cancer, 89*(4), 605-611. https://doi.org/10.1038/sj.bjc.6601120

[15] Ibeakuzie, P. O., & Onyeagu, S. I. (2024). A Parametric Cox Proportional Hazard Model with Application. *Earthline Journal of Mathematical Sciences, 14*(4), 747-771. https://doi.org/10.34198/ejms.14424.747771

[16] Duran-Struuck, R., & Dysko, R. C. (2009). Principles of bone marrow transplantation (BMT): providing optimal veterinary and husbandry care to irradiated mice in BMT studies. *Journal of the American Association for Laboratory Animal Science, 48*(1), 11-22.

[17] Urbano-Ispizua, A. (2007). Risk assessment in haematopoietic stem cell transplantation: stem cell source. *Best Practice & Research Clinical Haematology, 20*(2), 265-280. https://doi.org/10.1016/j.beha.2006.09.006

[18] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209-249. https://doi.org/10.3322/caac.21660

[19] West, N. P., Kobayashi, H., Takahashi, K., Perrakis, A., Weber, K., Hohenberger, W., Sugihara, K., & Quirke, P. (2012). Understanding optimal colonic cancer surgery: comparison of Japanese D3 resection and European complete mesocolic excision with central vascular ligation. *J. Clin. Oncol., 30*(15), 1763-1769. https://doi.org/10.1200/JCO.2011.38.3992

[20] Su, H., Xie, S., Wang, S., Huang, L., Lyu, J., & Pan, Y. (2024). New findings in prognostic factor assessment for adenocarcinoma of transverse colon: a comparison study between competing-risk and COX regression analysis. *Frontiers in Medicine, 11*. https://doi.org/10.3389/fmed.2024.1301487

[21] Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963-974. https://doi.org/10.2307/2529876

[22] Ibrahim, J. G., Chu, H., & Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology, 28*(16), 2796. https://doi.org/10.1200/JCO.2009.25.0654

[23] Asar, Ö., Ritchie, J., Kalra, P. A., & Diggle, P. J. (2015). Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. *International Journal of Epidemiology, 44*(1), 334-344. https://doi.org/10.1093/ije/dyu262

[24] Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics, 1*(4), 465-480. https://doi.org/10.1093/biostatistics/1.4.465

[25] Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press. https://doi.org/10.1201/b12208

[26] Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics 53*(1), 330-339. https://doi.org/10.2307/2533118