Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

PROJECT 2

Project Based on “Survival analysis of patients with spinal chordomas"

Sabrina Enriquez*

[1]Department of Mathematics, University of California, Davis. 1 Shields Ave., Davis, California , USA

Correspondence
*Sabrina Enriquez, Email: seenriquez@ucdavis.edu

Present Address
1 Shields Ave, Davis, CA 95616

Summary

We will be mimicking the SEER data used in “Survival analysis of patients with spinal chardomas" by Sun, Hong, Liu, Cui, Zhou, Xie, and Wu (Neurosurgical Review 2019)[1] using information about the data presented in the paper. The researchers in[1] studied the survival differences between patients who received radiotherapy (R), surgery (S), both radiotherapy and surgery (SR), and neither radiotherapy nor surgery (NRS) over a 40 year period. We will analyze our simulated data using methods learned in BST 222 and re-generate the major results found in the paper. We will conclude by discussing how our simulated data compares to the original SEER data used in the paper’s statistical analysis. Data simulation code is written in R and provided in the Appendix while all code for survival analysis is written and performed using SAS and also provided in the Appendix.

KEYWORDS:
survival, spinal chardoma, cox proportional hazard

1 | INTRODUCTION

Using survival analysis methods learned in BST 222 and implemented in[1], we will simulate the data used from the SEER data base and replicate their analysis. The objective of the paper we are discussing was “to analyze the survival of patients with spinal chordomas... using data in the Surveillance, Epidemiology and End Results (SEER) database"[1]. The authors were motivated to study spinal chordomas because they are rare malignant tumors and because “there is no previous study demonstrating the longitudinal changes in survival of different treatment groups over the past few decades"[1]. In the same spirit, here we attempt to reconstruct the SEER data using results presented in the paper to further investigate survival for patients with this rare and difficult form of cancer.

2 | BACKGROUND

The following definitions and concepts will be utilized throughout this project and all definitions are sourced from *Survival Analysis Techniques for Censored and Truncated Data* by Klein and Moeschberger[2] unless otherwise stated.

Definition 1. Survival Function: The basic quantity employed to describe time-to-event phenomena is the survival function, the probability of an individual surviving beyond time x (experiencing the event after time x). It is defined as

$$S(x) = Pr(X > x).$$

Definition 2. Hazard Function: This function is fundamental in survival analysis and is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, the inverse of the MillâĂŹs ratio in economics, or simply as the hazard rate. The hazard rate is defined by

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

If X is a continuous random variable, then,

$$h(x) = f(x)/S(x) = -d \log[S(x)]/dx.$$

Definition 3. Cumulative Hazard Function: The cumulative hazard function tells us the total risk that has been accumulated at time x and is defined by

$$H(x) = \int_0^x h(u)du = -\log[S(x)]$$

Definition 4. Kaplan- Meier estimator: The standard estimator of the survival function, proposed by Kaplan and Meier (1958), is called the Product-Limit estimator. This estimator is defined as follows for all values of t in the range where there is data:

$$\hat{S}(t) = \begin{cases} 1 & t \leq t_1 \\ \prod_{t_i \leq t} [1 - \frac{d_i}{Y_i}] & t_i \leq t \end{cases}$$

where d_i is number of deaths and Y_i is the number of individuals at risk at time t_i

Definition 5. Weibull distribution: The Weibull distribution is a continuous distribution characterized by a shape parameter $\alpha > 0$, and a scale parameter $\lambda > 0$ with probability density function for a Weibull random variable given by

$$\alpha \lambda x^{\alpha - 1} \exp(-\lambda x^{\alpha}).$$

The Weibull distribution is a very flexible model for lifetime data. It has a hazard rate which is either monotone increasing, decreasing, or constant. It is the only parametric regression model which has both a proportional hazards representation and an accelerated failure-time representation.

The survival function for the Weibull distribution is given by

$$S_X(x) = \exp(-\lambda x^{\alpha}),$$

and the hazard rate is expressed by

$$h_X(x) = \lambda \alpha x^{\alpha - 1}.$$

The likelihood function for right-censored data is given by

$$L = \prod_{j=1}^{n} [f_Y(y_j)]^{\delta_j} [S_Y(y_j)]^{(1-\delta_j)} \text{ s.t. } f_Y(y_j) = (1/\sigma) \exp[(y - \mu)/\sigma - \exp[(y - \mu)/\sigma]]$$

with δ_j indicating censorship.

Definition 6. Cox Proportional Hazards Model[3]: Cox(1972) proposed to model the hazard function by

$$h(t|\mathbf{Z}) = h_0(t) c(\beta^t \mathbf{Z}) = h_0(t) \exp(\sum \beta_k \mathbf{Z}_k)$$

where $h_0(t)$ is an arbitrary baseline hazard rate, $\beta = (\beta_1, ..., \beta_p)^t$ is a parametric vector and $c(\beta^t \mathbf{Z})$ is a known function. This is called a semi-parametric model because a parametric form is assumed only for the covariate effect and the baseline hazard rate is treated nonparametrically.

This model assumes

1. IID observations
2. Noninformative/independent censoring

3. Hazard ratio (HR) is independent of time
4. Hazard ratio for two Z's are proportional

The partial likelihood function is

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp[\sum_k \beta_k Z_{(i)k}]}{\sum_{j \in R(t_i)} \exp[\sum_k \beta_k Z_{jk}]}$$

where $R(t_i)$ is the risk set at time t_i which includes all individuals who are still under study at a time just prior to time t_i. Here we assume no ties between the event times and $Z_{(i)k}$ is the k-th covariate associated with the individual whose failure time is t_i.

(Definition credit: L. Qi lecture 12 slides)

3 | OVERVIEW OF RESEARCH OBJECTIVES AND RESULTS BY SUN ET AL

The authors used data from 765 cases diagnosed between 1974 and 2013 from the SEER database and conducted further analysis on a subset of 379 patients diagnosed between 2004 and 2013 in an effort to avoid bias caused by evolution in treatment. All of the patients either had tumors located on the *mobile spine* or *sacrum*. They were focused on investigating the influence of clinical factors on the overall and cancer-related survival of patients. Cancer-related or cancer-specific survival refers to the survival rates of patients with spinal chordomas with "classic" pathology (n=747) instead of "dedifferentiated" (n=6) and "chondroid"(n=12). The researchers truncated the data provided by SEER database (n=14) to exclude cases where the treatment strategy was unknown and proceeded to group the remaining 765 patients by treatment strategies and by year of diagnosis. Specifically, the cases were grouped according to which treatment methods they used between
(R): Radiotherapy only
(S): Surgery only
(SR): Surgery combined with Radiotherapy
(NRS): Neither Surgery nor Radiotherapy at all.

The authors chose to group by year of diagnosis according to the following four intervals: 1974-1983, 1984-1993, 1994-2003, and 2004-2013. These groupings were chosen to observe how evolution of therapies (which is dependent on time) might have contributed to survival. They provide the bar chart presented in Figure 1 showing the distribution of years of diagnosis.

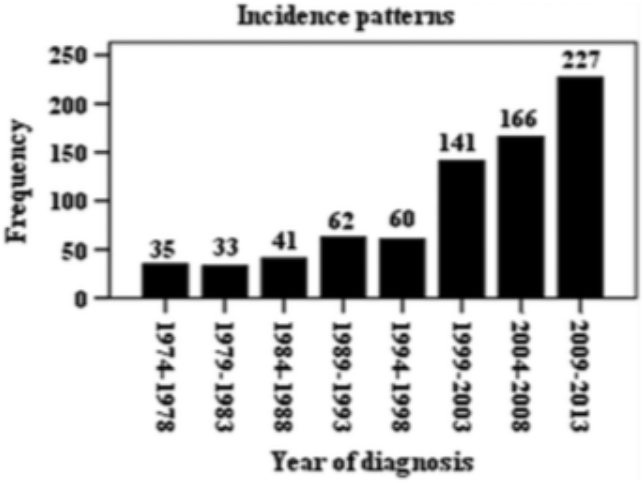


FIGURE 1 The incidence of spinal chordomas increased obviously during last few decades [1]

Using this structure, the demographics of patients presented in Figure 2, survival times, and characteristics of the cancer, the authors analyzed the changes in survival over time using Kaplan- Meier analysis and utilized Cox proportional hazards models for univariate and multivariate analysis to identify factors associated with analysis.

Table 1 Patient and tumor characteristics

Characteristics		Number (%)
Patients, *n*		765
Mean age (years) (± SD)		60.3 ± 17.1
Mean tumor size (mm) (± SD)		82.5 ± 74.2
Gender		
Men		475 (62.1)
Women		290 (37.9)
Marriage status	Married	459 (60.0)
	Never married	127 (16.6)
	Widowed	79 (10.3)
	Divorced	44 (5.8)
	Others	56 (7.3)
Race	White	676 (88.4)
	Asian/Pacific Islander	55 (7.2)
	Black	19 (2.5)
	Others	15 (2.0)
Tumor location	Mobile spine	339 (44.3)
	Sacrum	426 (55.7)
Pathology	Classic	747 (97.6)
	Chondroid	12 (1.6)
	Dedifferentiated	6 (0.8)

Data presented as number of patients (%) unless otherwise indicated

FIGURE 2 Demographics of SEER data[1]

3.1 | Survival outcome parameters

The following are the survival outcome parameters defined in this paper:
Event of interest: date of death.
Survival time origin: date of diagnosis.
Time scale: months since diagnosis.
Censoring: uninformative random right censoring at last follow-up.
Truncation: left truncation of patients with unknown treatment strategies. There were 14 patients excluded from this analysis.

3.2 | Primary questions and results

1. How do different treatment methods affect survival outcome for patients?

   (a) Outcomes are supported by Figure 3 and are summarized by:
   (R): significant differences (P= 0.037)
   (S): significant differences (P= 0.031)
   (SR): not significant improvement of survival (P=0.221)
   (NRS): significant differences (P= 0.031)

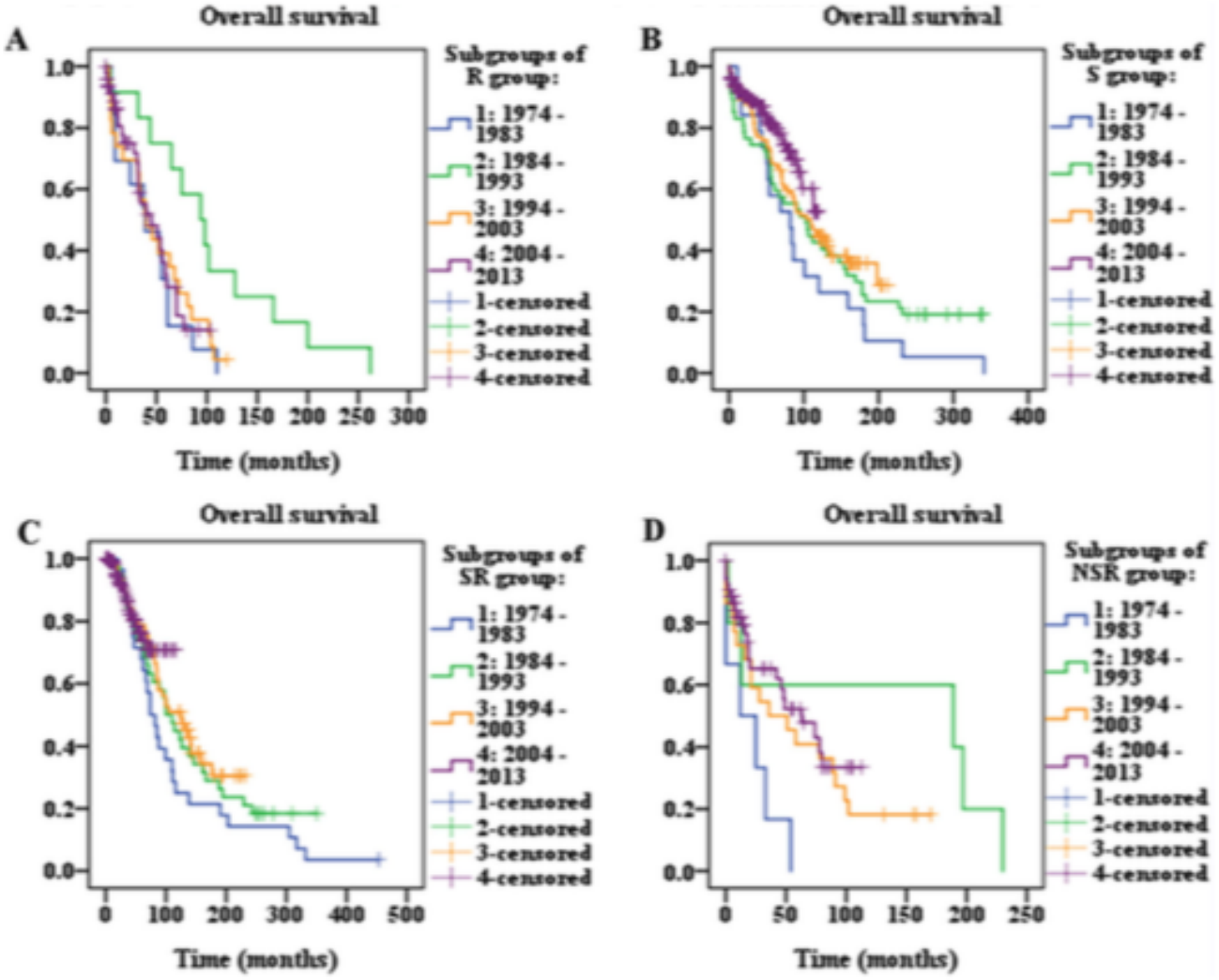


FIGURE 3 A: The difference in overall survival among subgroups of patients receiving R was statistically significant (P = 0.037), with patients in 1984 -1993 presenting the best overall survival. **B**: The difference in overall survival among subgroups of patients receiving S was statistically significant (P = 0.031), with survival rate increasing steadily over time. **C**: There is no significant difference in overall survival among subgroups of patients receiving SR (P = 0.221). **D**: The difference in overall survival among subgroups of patients receiving NSR was statistically significant (P = 0.031), with survival rate increasing steadily over time [1]

2. How does the site of the chordoma (skull base, mobile spine, and sacrum) affect survival?

   (a) Outcome: Sacrum is associated with better overall survival (HR 0.401, P=0.002). Sacrum location also associated with higher cancer-specific survival (HR 0.287, P=0.002). This result was further explained in Figures 3, 4, and 5.

3. How does type of chardoma (classic, chondroid, and dedifferentiated) affect survival?

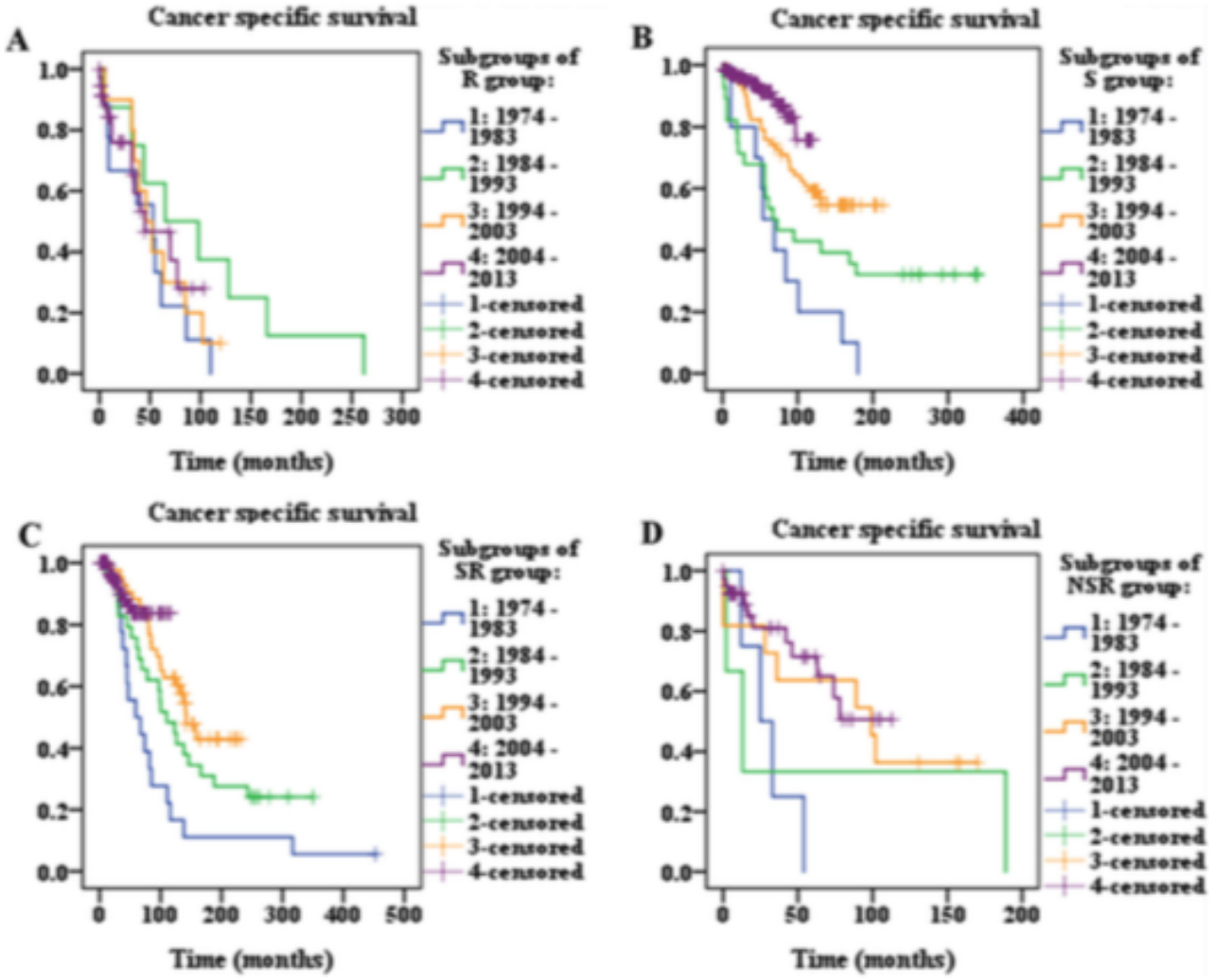


FIGURE 4 A: There is no significant difference in cancer-specific survival among subgroups of patients receiving R (P = 0.411). **B**: The difference in cancer- specific survival among subgroups of patients receiving S was statistically significant (P < 0.001), with survival rate increasing steadily over time. **C**: The difference in cancer-specific survival among subgroups of patients receiving SR was statistically significant (P = 0.001), with survival rate increasing steadily over time. **D**: The difference in cancer-specific survival among subgroups of patients receiving NSR was statistically significant (P = 0.049), with survival rate increasing steadily over time[1]

(a) Outcome: not significant with patients receiving R (P=0.411) and significant with patients receiving S, SR, NSR with respective (P<0.001, 0.001, and 0.049) as shown in 4. Sacrum location for these tumors was consistently better than mobile spine tumor location for overall and cancer-specific survival as shown in Figure 5.

4. How does age at time of diagnosis affect survival?

(a) Outcome: Cox regressional model showed younger onset age (hazard ratio [hr] 1.052, P<0.001) was associated with better overall survival and is also significant for cancer-specific survival (hazard ratio [hr] 1.036, P=0.029) as shown in Figures 6 and 7.

5. Has spinal chardoma patient's survival improved significantly over time?

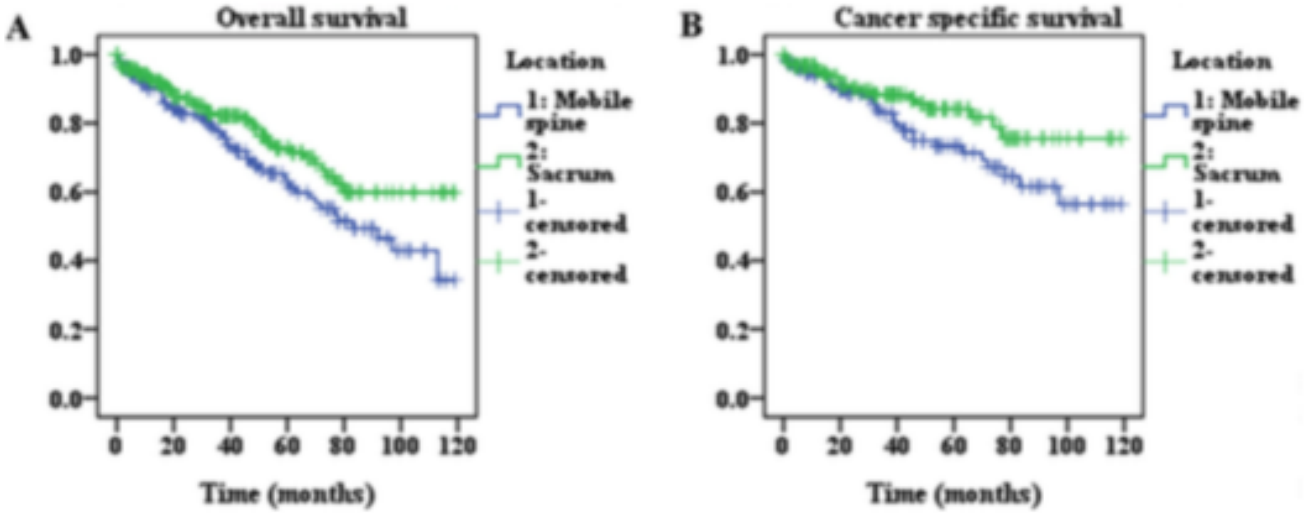


FIGURE 5 A: Patients with tumors on the sacrum presented significantly higher overall survival than those with tumors on mobile spine (P = 0.002). **B**: Patients with tumors on the sacrum presented significantly higher cancer-specific survival than those with tumors on mobile spine (P = 0.006) [1]

Table 2 Overall survival for patients with spinal chordomas: Cox proportional hazards analysis

	Univariate			Multivariate (final model)		
Variable	HR	95% CI	*P* value	HR	95% CI	*P* value
Onset age	1.052	1.036–1.068	< 0.001	1.052	1.027–1.078	< 0.001
Tumor size	1.001	0.999–1.002	0.525	–		
Married (vs others)	0.790	0.536–1.164	0.234	–		
Male (vs female)	1.138	0.762–1.699	0.528	–		
White race (vs others)	1.627	0.789–3.353	0.187	–		
Sacrum (vs mobile spine)	0.668	0.453–0.985	0.042	0.401	0.225–0.714	0.002
Surgery (vs no surgery)	0.288	0.195–0.424	< 0.001	0.291	0.139–0.610	0.001
Radiotherapy (vs no radiotherapy)	0.861	0.583–1.271	0.451	–		
Surgery and radiotherapy (vs others)	0.524	0.322–0.855	0.010	–		
Classic chordomas (vs dedifferentiated)	0.215	0.067–0.685	0.009	–		
Chondroid chordomas (vs dedifferentiated)	0.507	0.113–2.279	0.376	–		

No surgery means all other treatment options but surgery (R + SR). It is the same for the other treatment groups

FIGURE 6 Univariate and Multivariate Cox proportional hazards analysis for all 765 patients [1]

(a) Outcome: Incidence of spinal chardomas has increased over the last 40 years, but survival has improved steadily for patients receiving surgery. There was no significant difference in overall survival among patients receiving surgery combined with radiotherapy and surprisingly the patients receiving radiotherapy alone between 1984 and 1993 had the longest overall survival among subgroups defined by diagnosis date who received radiotherapy only. Lastly, patients receiving NSR (neither) also have improved survival over time.

6. How do clinical factors influence the overall and cancer- related survival for patients diagnosed between 2004 and 2013?

(a) Outcome:Those patients had the longest cancer-specific survival.

Table 3 Cancer-specific survival for patients with spine chordomas: Cox proportional hazards analysis

	Univariate			Multivariate (final model)		
Variable	HR	95% CI	*P* value	HR	95% CI	*P* value
Onset age	1.053	1.031–1.075	<0.001	1.036	1.004–1.070	0.029
Tumor size	1.001	0.999–1.003	0.420	–		
Married (vs others)	0.784	0.452–1.360	0.387	–		
Male (vs female)	1.070	0.609–1.880	0.813	–		
White race (vs others)	1.800	0.647–5.006	0.260	–		
Sacrum (vs mobile spine)	0.547	0.313–0.955	0.034	0.287	0.129–0.639	0.002
Surgery (vs no surgery)	0.208	0.120–0.362	<0.001	0.221	0.071–0.689	0.009
Radiotherapy (vs no radiotherapy)	0.655	0.377–1.138	0.133	–		
Surgery and radiotherapy (vs others)	0.541	0.278–1.056	0.072	–		
Classic chordomas (vs dedifferentiated)	0.086	0.020–0.374	0.001	–		
Chondroid chordomas (vs dedifferentiated)	0.219	0.030–1.607	0.219	–		

No surgery means all other treatment options but surgery. It is the same for the other treatment groups

FIGURE 7 Univariate and Multivariate Cox proportional hazards analysis for all 765 patients [1]

4 | DATA SIMULATION

The authors of this paper used a Cox proportional hazards model for the whole data set of 765 patients but then focused much of their analysis on a subset of patients diagnosed between 2004 and 2013. In that diagnosis time range there were 379 patients and the authors focused their analysis on them in an effort to avoid bias produced by the evolution of treatment strategies over time. While they provided many demographic statistics of the 765 patients they did not provide the specific demographics of the 379 patients they focused their analysis on. This causes an issue with mimicking the data, but we attempt to bridge this gap in information by first simulating the whole data set and using the distribution of diagnosis times we take the subset of patients diagnosed after 2004.

To simulate the data for the 765 patients we first turn our attention to the demographics presented in the paper and shown here in Figure 2. Using the framework our TA Xiner Zhou presented in Lab 4 [4], we simulate our data. First, we note that the the paper assumed the Cox PH model and did not specify the distribution of the baseline event time used in their model. Therefore we begin by assuming that the baseline event time follows a Weibull distribution and that there is noninformative random right censoring.

For simulating the demographic variables we use the distributions between groups specified in Figure 2 and generate 765 observations in accordance with those distributions for their corresponding 18 covariates. The first 2 covariates: age and tumor size are continuous variables so we use a random normal distribution to mutually independently simulate their values while the rest of the demographic covariates are generated using a random sample function in accordance with each subgroup's probability given by their respective percentages. This method is appropriate because it will ensure we simulate a set of 765 patients with demographics that match the original data's demographics rather closely, leaving room for small changes due to the random nature of the functions used. Using this method we generate characteristics Table 1 and rejoice that the simulated data's demographics reflect the original data's characteristics shown in Figure 2 closely.

4.1 | Ambiguities and omissions in the original paper

The following ambiguities or omissions in the original paper make near-perfect simulation impossible. To address these glaring problems we make assumptions and decide on specific interpretations of the author's writings to finish the simulation. We only make assumptions and judgement calls when absolutely necessary for moving forward with simulation.

TABLE 1 Patient and Tumor Characteristics

Characteristics		
Patients		765
Mean age (years) (+-SD)		60.37 +- 4.13
Mean tumor size (mm) (+-SD)		82.39+- 8.65
Gender	Frequency	Percent
Female	313	40.92
Male	452	59.08
Marital	Frequency	Percent
Married	461	60.26
Never Married	141	18.43
Widowed	69	9.02
Divorced	41	5.36
Other	53	6.93
Race	Frequency	Percent
White	693	90.59
Asian/ Pacific Islander	46	6.01
Black	16	2.09
Other	10	1.31
TumorLocation	Frequency	Percent
Mobile spine	327	42.75
Sacrum	438	57.25
Pathology	Frequency	Percent
Classic	740	96.73
Chondroid	18	2.35
Dedifferentiated	7	0.92

1. The authors omit the demographic descriptive statistics of the subset of patients diagnosed after 2004. To remedy we assume the distributions for the whole data set apply to the subset.

2. The authors are ambiguous about what “cancer-specific" group means and we infer that they mean patients with classic pathology.

3. The authors describe a significant factor as “young onset age" and give no value for the age used as a reference to mean “young" so we decide young means all ages before the reference value used in our multivariate Cox PH analysis.

4. The authors omit the distribution for how the patients are grouped between R, S, RS, and NSR for the 765 patients, but they do give it for the 379 patients diagnosed after 2004. To remedy we once again assume the distribution is consistent for the subset and the whole data set and use the distribution for the 379 for simulating all 765 patient’s treatment group distribution. This is necessary for reproducing their results for survival for all patients since the objective is to understand the influence of clinical factors. It is appropriate to do this because when we take the subset from the whole data set we can expect that the subset will have a distribution close to the observed data set’s distribution.

4.2 | Concluding the simulation given assumptions

After that process and making the aforementioned assumptions we use five adjusted hazard ratios with p-values less than 0.05 that were provided for the univariate Cox model shown in Figure6. By taking the negative natural logarithm of the hazard ratios we can find the β values corresponding to each of those five covariates: Age, sacrum, surgery, SR, and classic pathology. We only use these five hazard ratios because they are adjusted for the eleven covariates in their model which we are using to simulate our data. We also format our data to ensure we have each of the 11 covariates shown in their tables are accounted for. In their table shown in Figure 6 they indicate the reference values for each category and we code for that by assigning the non-reference as 1 in a binary dummy variable that is determined by the data we already simulated. Then, using our beta values and those dummy variables corresponding to young onset, sacrum, and surgery, we generate our survival time simulations using the random Weibull function with shape =1 and scale equal to

$$\lambda_0 \exp(\beta_1 * age + \beta_2 * sacrum + \beta_3 * surgery).$$

Then we use a right censor time = 480 after observing that to be the largest right censor time. Our arbitrary λ_0 is set to 200, because after testing different values we found that $\lambda_0 = 200$ distributed our time so that the median time is 52 months, which

matches the original data. Furthermore, we perform random uniform censoring, and using our function provided by Zhou[4], we conclude our data simulation procedure. The full simulation code written in R language is provided in the appendix.

5 | METHODS

In the original paper Kaplan- Meier survival analysis and Cox proportional hazard model were used for univariate and multivariate analysis to identify factors associated with survival for the whole data set. Then those same methods were applied to the subset of patients diagnosed after 2004 and adjusted hazard ratios were reported with 95% confidence intervals. We followed the same methods as the authors using our simulated data and in the final multivariate Cox model only 3 covariates were determined to be significant and therefore included. We include our SAS code in the appendix and compare our results to the originals.

After regenerating the major results from the original paper we conduct our own analysis to further investigate their findings. We utilize k-sample test techniques learned in BST 222 including the log-rank, Tarone, Peto and Fleming tests to see if we can find other associations between the four therapy types: R, S, SR, and NSR. Then we will plot the survival curves for each therapy and interpret our graphs. Finally we will do the opposite of Figure 3 and 4 by generating a panel of KM curves of therapy type over diagnosis groups. These methods are all appropriate to further understand the data because they shed light on how the individual therapies compare within each diagnosis time frame. This analysis will complement the author's findings since they were more focused on comparing the efficacy of therapies over time instead of reporting the survival rates within each period.

6 | RESULTS

As described in our methods, we will begin by regenerating the major results using our simulated data. First we check our simulated data and ensure that the distribution of diagnosis time is consistent with the observed data. It is crucial that we follow that distribution since we will be further investigating those diagnosed after 2004 as they did in the paper and we need roughly the same number of patients in each group if we expect similar results. Thus, we check the number of patients in each diagnosis group and present the distribution in Figure 8.

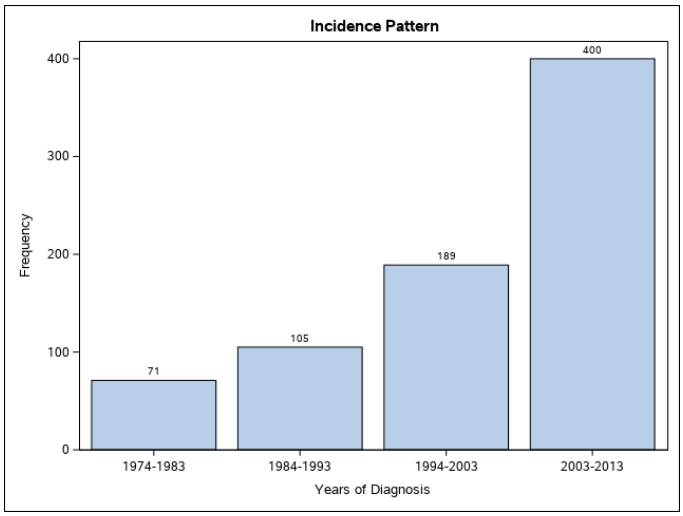


FIGURE 8 Incidence patterns of year of diagnosis for simulated data. Note that it closely matches the observed data's distribution shown in 1.

We remind the reader that in the original data shown in Figure 1 there were 68, 103, 201, and 393 patients in each diagnosis group in chronological order and in our simulation we have 71, 105, 189, and 400. Since we introduced randomness in our simulation these slight variations from the observed data and continue with reproducing their results. Our distribution is consistent with their observation of incidence increasing with time.

Next we create a panel of four KM curves showing how each diagnosis time group compares within the same therapy type in Figure 9. This figure is replicating the analysis done for Figure 3 and shows some of the ways that our simulated data varies from the observed SEER data. We have higher survival for patients diagnosed between 1974-1983 in the therapy =1 =Radiotherapy group while the original data has the second diagnosis group (1984-1993) as having the highest survival rate for radiotherapy. This difference is also explainable by our random assumption in simulation and also reflects that there are fewer patients from that period in that therapy group in our simulation than in the observed data. This may be a result of making assumptions about the distribution of therapies amongst all patients explained in 4.1.4. The rest of the therapies are rather consistent with the original with some added noise.

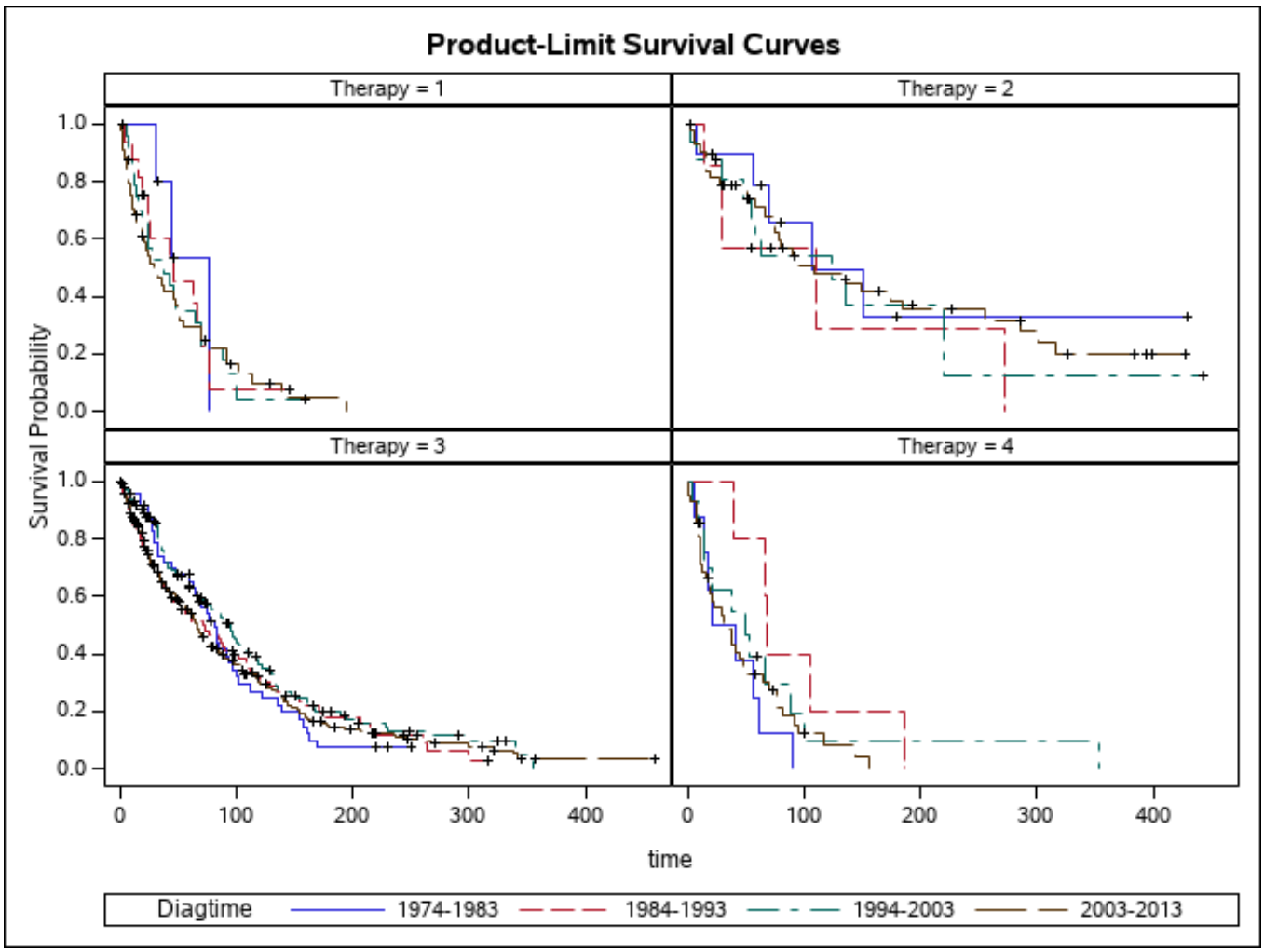


FIGURE 9 Kaplan- Meier plots analyzed over diagnosis year group for Therapy 1 = R, Therapy 2 = S, Therapy 3 = SR, and Therapy 4 = NSR of simulated data.

Next we analyze survival of diagnosis time groups over therapies restricted to cancer-specific cases. This panel is shown in Figure 10 and corresponds to Figure 4 in the original paper. As in the original paper our graphs are almost identical because 97.6% of the cancer pathology for all cases was classic and therefore very few cases were excluded in this analysis. In the original paper they include findings about how therapy type was associated with survival in the caption of this figure but do not explain the analysis methods they used for those findings nor show the KM curves that support their claims. We fill that gap in our extended analysis.

In Figure 11 we show survival curves of patients with mobile spine (in blue) versus sacrum tumor (in red) locations for all patients and for patients with classic pathology. Our KM curves are consistent with the curves shown in Figure 5. These

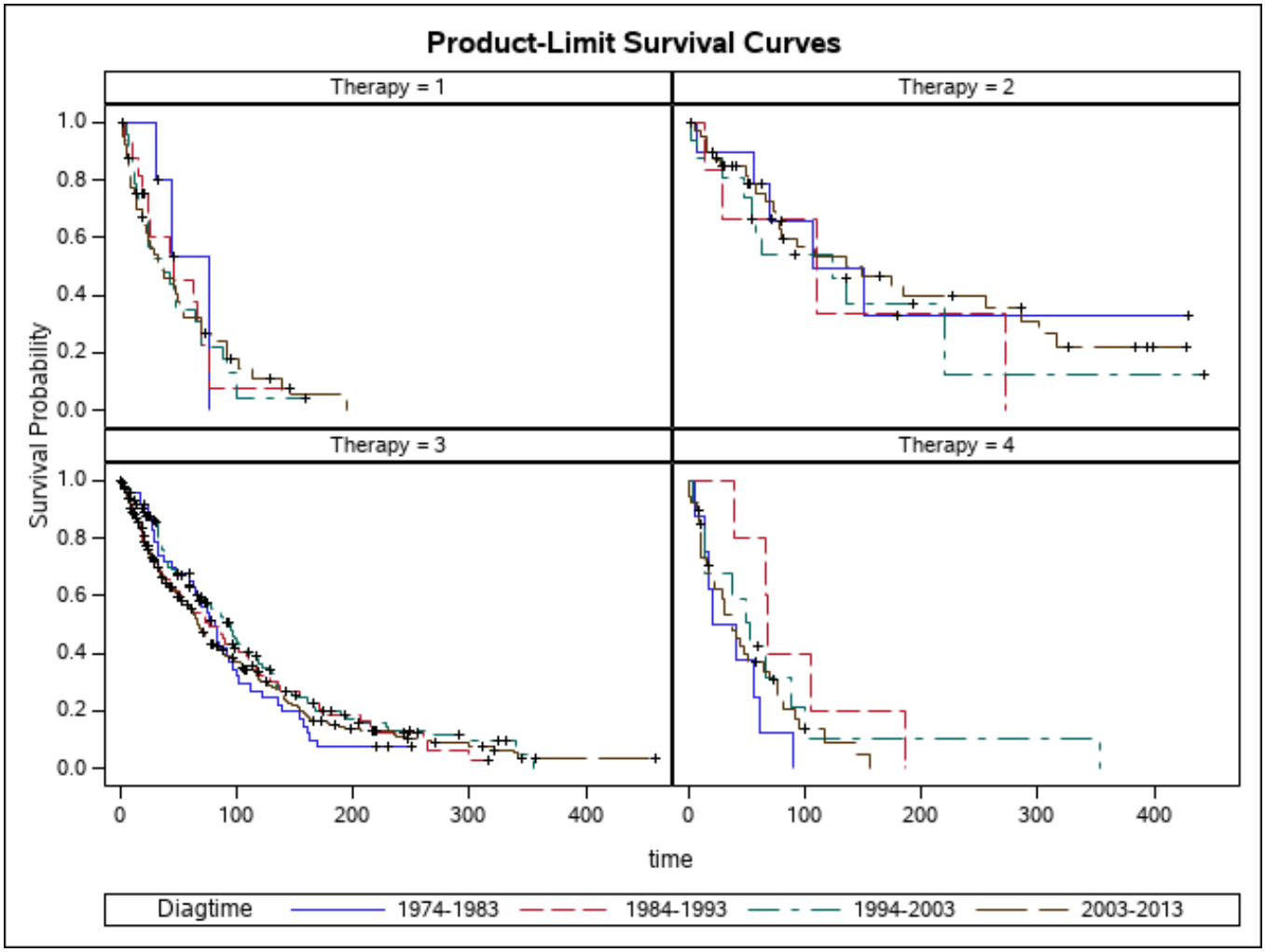


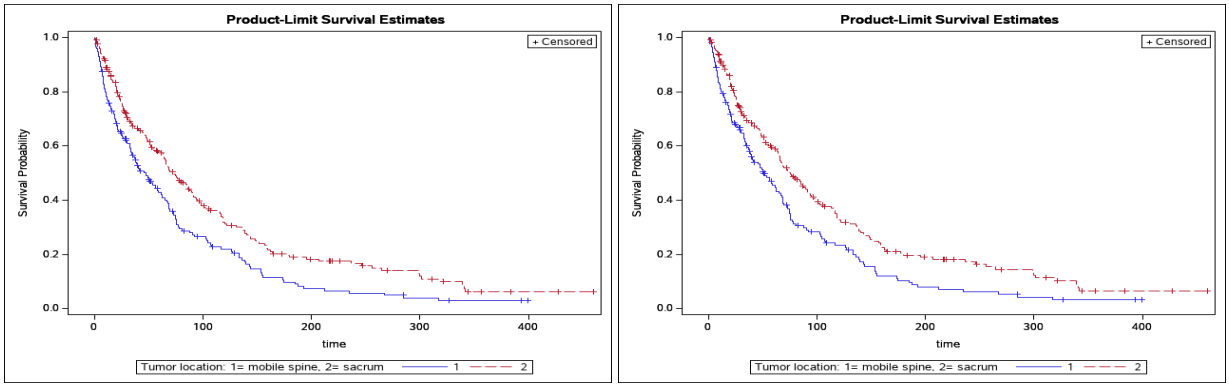
FIGURE 10 Cancer- specific Kaplan- Meier plots analyzed over diagnosis year group for Therapy 1 = R, Therapy 2 = S, Therapy 3 = SR, and Therapy 4 = NSR of simulated data.

KM curves show that sacrum tumor location has significantly better survival rate and we present the specific hazard rates and confidence bands in Table 4.

Lastly, we conclude comparing our results with the original paper's in Tables 4 and 5 found in the supporting information8 section. In each of those tables we report the adjusted hazard ratios for the cox univariate proportional hazards analysis for the 11 covariates outlined in Figures 6 and 7, as well as the adjusted hazard ratios for the final multivariate cox model. Our results are almost entirely consistent with those in the original paper which is expected since we used 5 of those observed hazard ratios to simulate our data. The only value that is more off from the original than anticipated are the values for chondroid chordomas. We found the hazard ratio to be 2.356 for overall survival and 2.981 for cancer- specific univariate analyses while the original paper reported 0.507 and 0.219 respectively. Our confidence bands are also much larger at 0.092-6.816 and 0.928-9.582 for overall and cancer specific respectively. Considering there are so few cases falling into that category it makes sense that our results can not match the original. In the original model it was not found to be significant and with only 12 cases total and likely fewer in the subgroup considered in this analysis we accept that our results are not precise for this covariate.

6.1 | Extended analysis

Here we conduct our own analysis and present interpretations of our results. First we perform Kaplan Meier analysis on the four therapy types R, S, SR, and NSR and show the survival curve in Figure 12.



(a) Overall Survival

(b) Cancer- Specific Survival

FIGURE 11 Kaplan- Meier curves showing that sacrum (red) maintains higher overall and cancer-specific survival than mobile spine. These plots align with the results shown in Figure 5. Here we get higher overall survival for sacrum location over mobile spine (P=0.0004) and higher cancer specific survival for sacrum location over mobile spine (P=0.0009) given by log rank test. Note that the plots shown in Figure 5 are cropped to only the first 120 months while these plots are not cropped and therefore might look different on first glance.

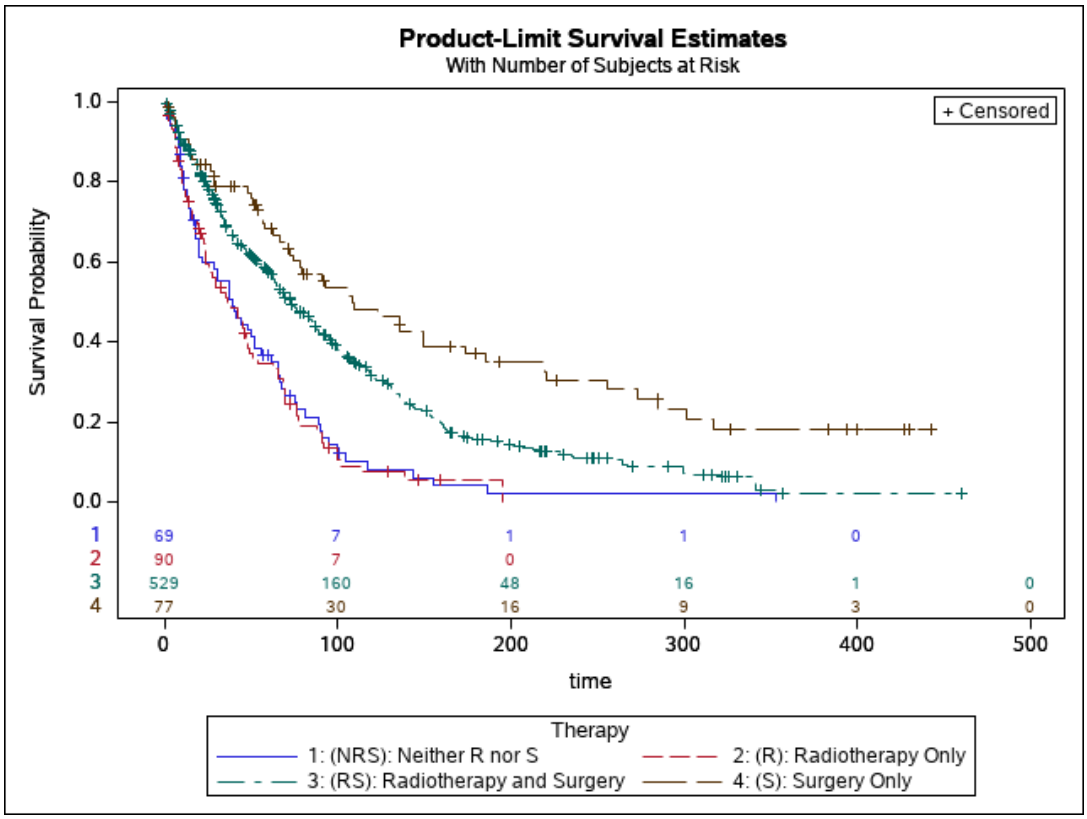


FIGURE 12 KM curves of therapy types for all patients.

We then perform Log-Rank, Tarone, Peto, Modified Peto, and Fleming(0,1) tests on therapy type shown in Table 2 to determine significance for survival outcome and find that therapy type is a significant factor.

After confirming the significance of therapy type we perform trend tests and present those results in table 3. Combining these results we are able to confirm that survival outcome is best for surgery only and better for radiotherapy and surgery, but radiotherapy only and neither are rather indistinguishable. This result is consistent with S and SR having significant p-values

TABLE 2 Rank statistics for log-rank, Tarone, Peto, Modified Peto, and Fleming(0,1) tests on therapy type for all patients. The tests show that therapy type is significant for survival outcome.

Rank Statistics					
Therapy	Log-Rank	Tarone	Peto	ModifiedPeto	Fleming
(NRS): Neither R nor S	26.189	547	16.996	16.957	9.168
(R): Radiotherapy Only	34.891	744	23.298	23.247	11.56
(RS): Radiotherapy and Surgery	-24.186	-694	-23.424	-23.4	-0.736
(S): Surgery Only	-36.895	-597	-16.87	-16.803	-19.992
Test of Equality over Strata					
Test	Chi-Square	DF	Pr >Chi-Square		
Log-Rank	67.978	3	<.0001		
Tarone	60.1971	3	<.0001		
Peto	54.8425	3	<.0001		
Modified Peto	54.7805	3	<.0001		
Fleming(0,1)	53.7416	3	<.0001		

in the original paper's univariate cox analysis. However, radiotherapy did not make it into the final multivariate model and graphically that makes sense since it is almost exactly in between the best therapy: S and the worst: NSR and R. Moreover, it does not satisfy the PH assumption since it intersects NRS around 350 months.

TABLE 3 Trend tests over therapy type.

Trend Tests						
Test	**TestStatistic**	**Standard Error**	**z-Score**	**Pr >|z|**	**Pr <z**	**Pr >z**
Log-Rank	-124.1648	16.1273	-7.6991	<.0001	<.0001	1.0000
Tarone	-2435.2973	335.8789	-7.2505	<.0001	<.0001	1.0000
Peto	-74.1589	10.7473	-6.9002	<.0001	<.0001	1.0000
Modified Peto	-73.9632	10.7261	-6.8956	<.0001	<.0001	1.0000
Fleming(0,1)	-49.8866	7.4741	-6.6746	<.0001	<.0001	1.0000

To conclude this extended analysis we flip the strata and group used to analyze survival in Figures 3 and 4 of the original study to better understand how therapy type was associated with survival within each diagnosis group period. In Figure 13 we show KM survival curves of therapy types for all patients in each diagnosis time group. These plots show that as time progresses RS remains consistently in the middle, R is consistently the worst option, and S and NRS vary more. Surgery seems to be the best option in 1974-1983 and 2004-2013, but in between there is much more variation and has survival outcomes most similar to RS. We also note that in the first and fourth time periods the therapy survival curves are much more distinguished from one another indicating more efficacy disparities between the methods during those times. In the final time period the S, RS, and NRS or S, RS, and R satisfy the PH assumption. These curves shed more light on why it was approrpiate to use cox models for only this time period instead of attempting to fit the cox PH model for all patients in the data.

For our last point of analysis we run the same analysis as before but restrict to cancer-specific patients with pathology type classic. This panel of KM curves is shown in Figure 14. As with the original paper so few cases are excluded here and therefore this analysis is equivalent to the analysis for all patients. We find the same conclusions here, but include it to finish filling in the gaps found in the original analysis.

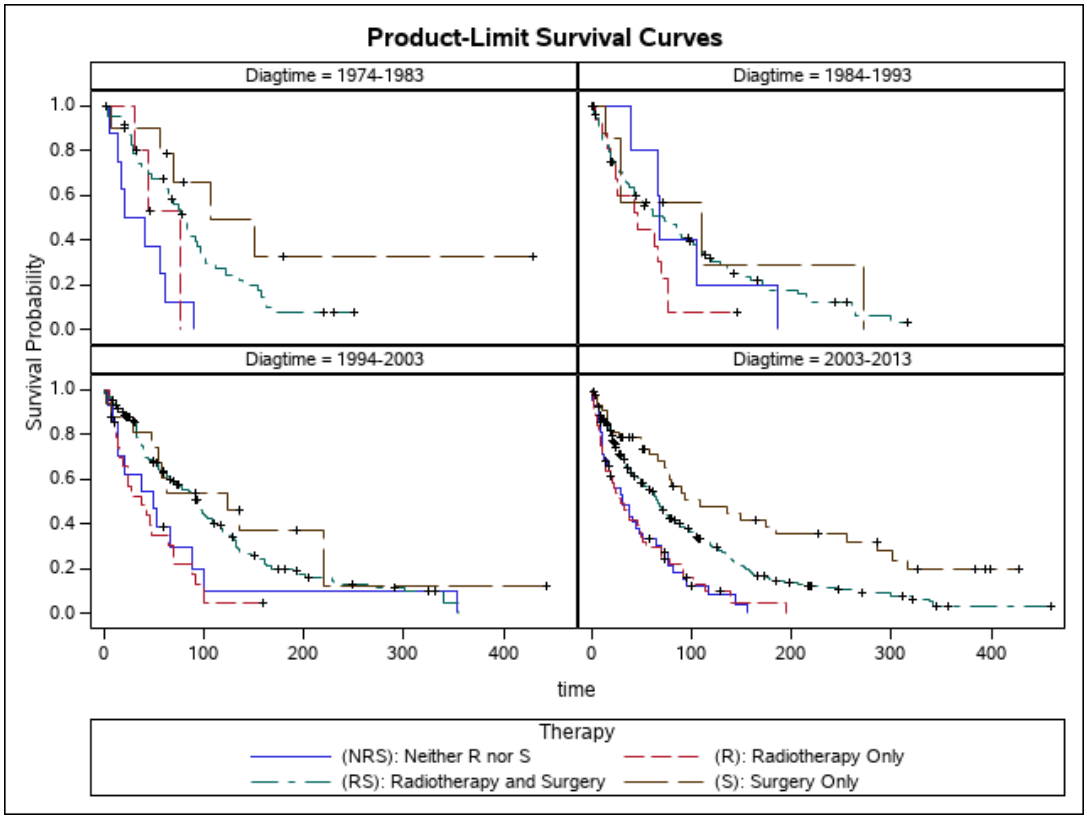


FIGURE 13 KM curves of therapy types for all patients stratified by diagnosis time periods.

7 | DISCUSSION

My simulated data mirrored the original data with added noise from the introduction of random variables. As explained in section 4.1 there were many limitations to reproducing this data any closer. The authors of the original data could improve our ability to simulate this data by including the information described in 4.1.1-4.1.4. Most notably this simulation could have had much closer results if demographic descriptions for the patients diagnosed after 2004 were provided as well as the distribution of treatment type for all 765 patients. The assumptions we were forced to make during simulation lead to inevitable noise but despite these setbacks we were still able to replicate the study effectively and our results were largely consistent with those presented in the paper *Survival analysis of patients with spinal chordomas*[1].

8 | CONCLUSION

In conclusion, via data simulation using the Cox Proportional Hazards Regression Model we were able to reconstruct SEER data and replicate the survival analysis done in *Survival analysis of patients with spinal chordomas*[1]. We extended their analysis to focus on therapy type and found that along with their conclusions, surgery then radiotherapy combined with surgery are the first and second best options for patients with spinal chordomas.

SUPPORTING INFORMATION

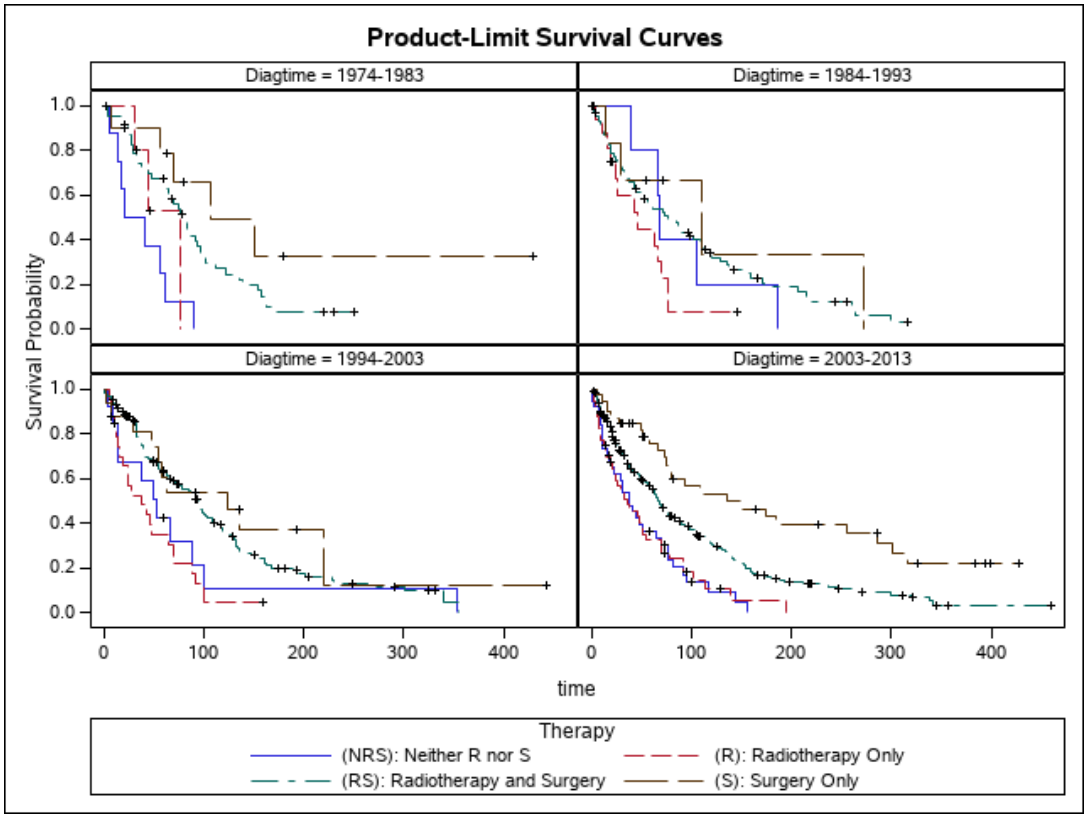


FIGURE 14 Cancer- specific KM curves of therapy types for all patients stratified by diagnosis time periods.

TABLE 4 Overall Survival for simulated patients diagnosed between 2004-2013: Cox proportional hazards analysis.

	Univariate				Multivariate (final model)			
Variable	HR	95% CI		P-Value	HR	95% CI		P-Value
Age	1.058	1.027	1.089	0.0002	1.047	1.018	1.077	0.0014
TumorSize	1.008	0.995	1.021	0.2404				
Married	1.149	0.909	1.453	0.2452				
Male	0.899	0.71	1.138	0.3753				
White	1.101	0.758	1.597	0.614				
Sacrum	0.665	0.524	0.842	0.0007	0.608	0.485	0.763	<.0001
Surgery	0.274	0.163	0.46	<.0001	0.432	0.29	0.644	<.0001
Radiotherapy	0.98	0.616	1.56	0.9329				
Surgery and Radiotherapy	0.559	0.383	0.816	0.0026				
Classic	0.231	0.815	0.582	0.1137				
Chondroid	2.356	0.092	6.816	0.0019				

TABLE 5 Cancer-specific survival for simulated patients diagnosed between 2004-2013: Cox proportional hazards analysis.

	Univariate				Multivariate (final model)			
Variable	HR	95% CI		P-Value	HR	95% CI		P-Value
Age	1.052	1.021	1.084	0.0008	1.044	1.014	1.075	0.0037
TumorSize	1.009	0.996	1.022	0.1902				
Married	1.1	0.866	1.397	0.4369				
Male	0.881	0.691	1.123	0.3058				
White	1.154	0.787	1.693	0.4637				
Sacrum	0.655	0.513	0.836	0.0007	0.607	0.48	0.766	<.0001
Surgery	0.233	0.134	0.405	<.0001	0.385	0.251	0.591	<.0001
Radiotherapy	0.88	0.539	1.436	0.609				
Surgery and Radiotherapy	0.512	0.346	0.759	0.0009				
Classic	0.325	0.134	0.791	0.0133				
Chondroid	2.981	0.928	9.582	0.0667				

References

1. Sun HH, Hong X, Liu B, et al. Survival analysis of patients with spinal chordomas. *Neurosurgical Review* 2019; 42(2): 455–462.

2. Klein JP, Moeschberger ML. *Survival Analysis Techniques for Censored and Truncated Data*. second ed. 2003.

3. Qi L. STA/BST 222 Survival Analysis Lecture 12. 2020.

4. Zhou X. Lab 4: Simulation.R. 2020.

APPENDIX

finalsim.R

sabrina

2020-12-18

```
#########################################################################
# simulation if cox PH model is assumed, with some continuous covariates;
# baseline event time is assumed to follow Weibull/exponential distribution
# indepedent (uniform) censoring and Right censoring
#########################################################################
library(MASS)


sim_cox<- function(N,lambda0, beta, censor.right)
{
  # N = Total sample size
  # beta = PH coefficients
  # lambda0 = rate parameter of the exponential distribution for baseline


  #gender=0 is female and gender =1 is male
  gender <- sample(x=c(0, 1), size=N, replace=TRUE, prob=c(0.379, 0.621))
  #marriage=1: married, 2: never married, 3:Widowed, 4:divorced, 5:others
  marital<-sample(x=c( 1, 2, 3, 4, 5), size=N, replace=TRUE, prob=c(0.6, 0.166, 0.103, 0.058, 0.073))
  #race =1: white, 2: API, 3: black, 4: others
  race<-sample(x=c( 1, 2, 3, 4), size=N, replace=TRUE, prob=c(0.884, 0.072, 0.025, 0.02))
  #tumorLoc= 1: Mobile spine, 2:sacrum
  tumorLoc<-sample(x=c( 1, 2), size=N, replace=TRUE, prob=c(0.443, 0.557))
  # #patho =1: classic, 2: chondroid, 3: dedifferentiated
  patho<-sample(x=c( 1, 2, 3), size=N, replace=TRUE, prob=c(0.976, 0.016, 0.008))
  # #diagtime =1 1974-1983 =2: 1984-1993 =3: 1994-2003 =4: 2004-2013
  diagtime<-sample(x=c( 1, 2, 3, 4), size=N, replace=TRUE, prob=c(0.089, 0.134, 0.262, 0.515))
  # #therapy =1: radiology =2:Surgery alone =3: Surgery and radiotherapy alone =4: neither surgery nor
  therapy<-sample(x=c( 1, 2, 3, 4), size=N, replace=TRUE, prob=c(0.1, 0.1, 0.7, 0.1))

  # generate continuous covariates, mutually indepedent
  #going to assume zero correlation between the two

  X = mvrnorm(N,mu=c(60.3,82.5),Sigma=matrix(c(17.1,0,0,74.2),2,2))
  age=X[,1]
  tumorSize=X[,2]


  # initial data set
  initial<-data.frame(id=1:N,
                      Gender=gender,
                      Marital = marital,
```

```
                    Race= race,
                    TumorLocation= tumorLoc,
                    Pathology=patho,
                    Diagtime= diagtime,
                    Therapy= therapy,
                    Age=age,
                    TumorSize=tumorSize)


#sacrum location =1 if tumorloc = 2
sacrum<- c(1:N)*0;
for (i in 1:N) {
  if(initial$TumorLocation[i]==2)
  {
    sacrum[i]<- 1;
  }
}

#surgery  =1 if therapy = 2
surgery<- c(1:N)*0;
for (i in 1:N) {
  if(initial$Therapy[i]==2)
  {
    surgery[i]<- 1;
  }
}

#radiotherapy  =1 if therapy = 1
rad<- c(1:N)*0;
for (i in 1:N) {
  if(initial$Therapy[i]==1)
  {
    rad[i]<- 1;
  }
}

#RS  =1 if therapy = 3
RS<- c(1:N)*0;
for (i in 1:N) {
  if(initial$Therapy[i]==3)
  {
    RS[i]<- 1;
  }
}

#classic  =1 if pathology = 1
classic<- c(1:N)*0;
for (i in 1:N) {
  if(initial$Pathology[i]==1)
  {
    classic[i]<- 1;
  }
}
```

```
  #chondroid  =1 if pathology=2
  chon<- c(1:N)*0;
  for (i in 1:N) {
    if(initial$Pathology[i]==2)
    {
      chon[i]<- 1;
    }
  }

  initial<-cbind(initial, sacrum, surgery, rad, RS, classic, chon)

  # generate underlying event time
  # T <- rweibull(n=N, shape=1, scale = lambda0*exp(beta[1]*young+beta[2]*surgery+beta[3]*sacrum))
  #

  T <- rweibull(n=N, shape=1, scale = lambda0*exp(beta[1]*age+beta[2]*sacrum
                                                  +beta[3]*surgery + beta[4]*RS + beta[5]*classic))

  #mean(X)
  #rexp(n=N, rate=lambda0*exp(beta*A))

  # censoring times
  ctime = runif(N, min=0, max=censor.right)

  # follow-up times and event indicators
  # time= c(1:N)*0
  # for(i in 1:N)
  # { if(initial$Diagtime[i]==4)
  # {
  #   time[i]<- pmin(T, ctime, 160)
  # }
  #   else
  #   {
  #     time[i] <- pmin(T, ctime, censor.right)
  #   }
  #
  # }

  time<- pmin(T, ctime, censor.right)

  censor <- as.numeric(T>ctime | T>censor.right)
  finalData<-cbind(initial, time, censor)

  return(finalData)

}
#median follow up time was 52 months so for lambda0=200 we get median time approx 52.
#latest censor time was 480 months so censor.right=480

finalSimP2<-sim_cox(N=765, lambda0=200, beta=c(-log(1.052), -log(0.668),-log(0.288), -log(0.524),-log(0

#data check
median(finalSimP2$time)
```

```
## [1] 53.11406
```

```
mean(finalSimP2$Age)
```

```
## [1] 60.36272
```

```
mean(finalSimP2$TumorSize)
```

```
## [1] 82.62102
```

```
sum(finalSimP2$Gender==0)
```

```
## [1] 287
```

```
sum(finalSimP2$time< 60)
```

```
## [1] 412
```

```
sum(finalSimP2$time< 120)
```

```
## [1] 592
```

```
sum(finalSimP2$Diagtime==4)
```

```
## [1] 389
```

```
write.csv(finalSimP2,
          file="/Users/sabrina/Desktop/Fall 2020/BST222/Project 2/finalsimp2.csv", row.names = FALSE)
```

```
/*Sabrina Enriquez project 2*/
LIBNAME p2 "/folders/myfolders/p2";


PROC import DATAFILE= "/folders/myfolders/p2/finalsimp2.csv"
    DBMS=csv
     out=p2.sim
        replace;

    GETNAMES=YES; /*Option GETNAMES determines whether to generate SAS variable names from the data values in the f
                   record of the imported file.*/
RUN;

data p2.years;
set p2.sim;
keep Diagtime;
run;
PROC FORMAT;
  VALUE  Diagnosistime 1="1974-1983"
                       2 ="1984-1993"
                       3  ="1994-2003"
                           4="2003-2013";

RUN;

PROC FORMAT;
  VALUE  therapy 1="(R): Radiotherapy Only"
                       2 ="(S): Surgery Only"
                       3  ="(RS): Radiotherapy and Surgery"
                           4="(NRS): Neither R nor S";

RUN;
*fig1;
/* ods listing gpath= '/folders/myfolders/p2/plotsandfigs'; */
/* ods graphics / imagename= "fig1" imagefmt=png; */
title 'Incidence Pattern';
PROC SGPLOT DATA=p2.years;
   FORMAT  Diagtime Diagnosistime.;
   Vbar Diagtime / datalabel;
      label Diagtime= 'Years of Diagnosis';
RUN;


*fig2;
/* ods listing gpath= '/folders/myfolders/p2/plotsandfigs'; */
/* ods graphics / imagename= "fig2" imagefmt=png; */
proc lifetest data=p2.sim
plots=survival(strata=panel) ;
FORMAT  Diagtime Diagnosistime.;
*FORMAT  Therapy therapy.;
time time*censor(1); strata Therapy/ group= Diagtime; run;

/* ODS TAGSETS.csv */
/* file= '/folders/myfolders/p2/plotsandfigs/therapycox.csv' */
/* style= minimal; */
proc phreg
data= p2.sim
plots(overlay) = survival;
FORMAT  Therapy therapy.;
class Therapy;
model time*censor(1) = Therapy;
run;
/* ods tagsets.csv close; */

*fig3;
*cancer specific to classic;
data p2.cspec;
set p2.sim;
if classic=1;
run;


/* ods listing gpath= '/folders/myfolders/p2/plotsandfigs'; */
/* ods graphics / imagename= "fig3" imagefmt=png; */
proc lifetest data=p2.cspec
plots=survival(strata=panel) ;
FORMAT  Diagtime Diagnosistime.;
time time*censor(1); strata Therapy/ group= Diagtime; run;

*fig4a
```

```
*overall survival by tumor location;
*we need to take the subset of data with diagtime ==4;
data p2.d4;
set p2.sim;
if DiagTime=4;
run;

data p2.cspecd4;
set p2.d4;
if classic=1;
run;



/* ods listing gpath= '/folders/myfolders/p2/plotsandfigs'; */
/* ods graphics / imagename= "fig4a" imagefmt=png; */
title 'Overall Survival by tumor location';
proc lifetest data=p2.d4
plots=survival ;
time time*censor(1);
strata TumorLocation;
label TumorLocation= 'Tumor location: 1= mobile spine, 2= sacrum';
run;
*fig4b
*cancer specific survival for classic;
/* ods listing gpath= '/folders/myfolders/p2/plotsandfigs'; */
/* ods graphics / imagename= "fig4b" imagefmt=png; */
proc lifetest data=p2.cspecd4
plots=survival ;
time time*censor(1);
strata TumorLocation;
title 'Cancer Specific Survival';
label TumorLocation= 'Tumor location: 1= mobile spine, 2= sacrum';
run;


PROC FORMAT;
  VALUE  gender 1="Male"
                0 ="Female";
    Value marriage 1="Married"
                2 ="Never Married"
                3= "Widowed"
                4="Divorced"
                5="Other";
    Value race 1="White"
                2 ="Asian/ Pacific Islander"
                3= "Black"
                4="Other";
    Value loc 1="Mobile spine"
           2= "Sacrum";
    Value path 1="Classic"
               2="Chondroid"
               3="Dedifferentiated";


RUN;


/* ODS TAGSETS.csv */
/* file= '/folders/myfolders/p2/plotsandfigs/table1.csv' */
/* style= minimal; */
proc means
data=p2.sim;
var Age TumorSize;
run;

PROC FREQ DATA=p2.sim;
   FORMAT  Gender gender.
           Marital    marriage.
           Race race.
           TumorLocation loc.
           Pathology path.;
   TABLES Gender Marital Race TumorLocation Pathology;
RUN;
/* ods tagsets.csv close; */



*cancer specific to chondroid;
data p2.chon;
set p2.sim;
```

```
if chon=1;
run;

proc lifetest data=p2.chon
plots=survival(strata=panel) ;
time time*censor(1); strata Therapy/ group= Diagtime; run;

*for this I need dummy variables for married vs others, white race vs others,
surgery vs others, rad vs others, RS vs others, classic vs dediff, chondroid vs dediff.
We made all of these except for white race and married so we do that now and add 2 columns.;

*married=1 vs others=0;
data p2.married;
set p2.d4;
if Maritial=1 then married = 1;
else married = 0;
run;


*white=1 vs others=0;
data p2.white;
set p2.married;
if Race=1 then white = 1;
else white = 0;
run;

*now i want to produce the hr for table 2;
/* ODS TAGSETS.csv */
/* file= '/folders/myfolders/p2/plotsandfigs/table2.csv' */
/* style= minimal; */
proc phreg data= p2.white plots(overlay) = (survival);
FORMAT  Gender gender.
           TumorLocation loc.
           Therapy therapy.
           Pathology path.;
class married(desc) Gender(desc) white(desc) TumorLocation(desc) surgery(desc) rad(desc) RS(desc)  Pathology;
model time*censor(1) = Age TumorSize married Gender white TumorLocation surgery rad RS Pathology/
Ties= EXACT RISKLIMITS ALPHA=.05;
run;
/* ods tagsets.csv close; */

*now the multivariate model;
/* ODS TAGSETS.csv */
/* file= '/folders/myfolders/p2/plotsandfigs/table2b.csv' */
/* style= minimal; */
proc phreg data= p2.white plots(overlay) = (survival);
FORMAT     TumorLocation loc.;
class  TumorLocation(desc) surgery(desc) ;
model time*censor(1) = Age TumorLocation surgery /
Ties= EXACT RISKLIMITS ALPHA=.05;
run;
/* ods tagsets.csv close; */


*let's make table 3;
*cancer specific to classic;
data p2.cspecd4;
set p2.white;
if classic=1 ;
run;




*cancer specific to dediff;
data p2.dediff;
set p2.white;
if Pathology=3;
run;
*cancer specific to classic;
data p2.chon2;
set p2.white;
if chon=1 ;
run;

*dediff and classic;
data p2.dc;
set p2.cspecd4 p2.dediff;

*chon and classic;
```

```
data p2.cc;
set p2.chon2 p2.dediff;
run;



*generate table 3;
/* ODS TAGSETS.csv */
/* file= '/folders/myfolders/p2/plotsandfigs/table3.csv'; */
proc phreg data= p2.cspecd4 ;
FORMAT  Gender gender.
          TumorLocation loc.

          Pathology path.;
class married(desc) Gender(desc) white(desc) TumorLocation(desc) surgery(desc) rad(desc) RS(desc) ;
model time*censor(1) = Age TumorSize married Gender white TumorLocation surgery rad RS /
Ties= EXACT RISKLIMITS ALPHA=.05;run;

proc phreg data= p2.dc ;

class classic(desc) ;
model time*censor(1) = classic /
Ties= EXACT RISKLIMITS ALPHA=.05;

proc phreg data= p2.cc plots(overlay) = (survival);

class chon(desc) ;
model time*censor(1) = chon /
Ties= EXACT RISKLIMITS ALPHA=.05;

run;
/* ods tagsets.csv close;  */

*multivariate;
/* ODS TAGSETS.csv */
/* file= '/folders/myfolders/p2/plotsandfigs/table3b.csv'; */
proc phreg data= p2.cspecd4 ;
FORMAT  Gender gender.
          TumorLocation loc.

          Pathology path.;
class TumorLocation(desc) surgery(desc);
model time*censor(1) = Age TumorLocation surgery /
Ties= EXACT RISKLIMITS ALPHA=.05;run;
/* ods tagsets.csv close;  */
*that's it for regenerating the results;







*now we do trend tests for the full set;

*K- sample test;
/* ods listing gpath= '/folders/myfolders/p2/plotsandfigs'; */
/* ods graphics / imagename= "t1" imagefmt=png; */
ODS TAGSETS.csv
file= '/folders/myfolders/p2/plotsandfigs/t1.csv';
title 'KM survival curves by treatment type for all patients';
proc lifetest data=p2.sim
plots=(survival (atrisk= 0 to 765 by 100)) ;
Format Therapy therapy.;
time time*censor(1); strata Therapy /
test= (logrank tarone peto modpeto fleming(0,1) ) ;
run;



*now trend test;
ods listing gpath= '/folders/myfolders/p2/plotsandfigs';
ods graphics / imagename= "t2" imagefmt=png;
proc lifetest data=p2.sim
plots=survival(atrisk= 0 to 765 by 100) ;
Format Therapy therapy.;
time time*censor(1); strata Therapy /
trend test= (logrank tarone peto modpeto fleming(0,1) ) ;
run;
/* ods tagsets.csv close; */
```

```
*now do opposite of their plots;
ods listing gpath= '/folders/myfolders/p2/plotsandfigs';
ods graphics / imagename= "t3" imagefmt=png;
proc lifetest data=p2.sim
plots=survival(strata=panel) ;
FORMAT  Diagtime Diagnosistime.;
FORMAT  Therapy therapy.;
time time*censor(1); strata Diagtime/ group= Therapy; run;


ods listing gpath= '/folders/myfolders/p2/plotsandfigs';
ods graphics / imagename= "t4" imagefmt=png;
proc lifetest data=p2.cspec
plots=survival(strata=panel) ;
FORMAT  Diagtime Diagnosistime.;
FORMAT  Therapy therapy.;
time time*censor(1); strata Diagtime/ group= Therapy; run;
```