SURVIVAL ANALYSIS OF PATIENTS WITH ORAL CANCER

SHUYI TAN

1. Introduction

According to the Oral Cancer Foundation, oral cancer is a particularly dangerous disease in its early stages it may not be noticed by the patient without producing pain or symptoms they might readily recognize. With the critical development of oral cancer screening as integral part of a clinician's routine, the disease is usually diagnosed at an advanced stage. In this study, the survival outcomes of 338 patients who were diagnosed with oral squamous cell carcinoma (OSCC) in the northernmost province of Finland between January 1, 1985 and December 31, 2005 was examined along with their demographic characteristics. To evaluate the association between survival rates and various prognostic factors in OSCC patients, both non-parametric and semi-parametric methods/models will be employed to fufill the analysis in this project.

2. Data Description and Exploratory Data Analysis

2.1. **Dataset Overview.** The dataset 'ORALCA' contains clinical records of 338 patients from a population-based retrospective cohort design, in which patients were diagnosed with oral squamous cell carcinoma (OSCC) in the northernmost province of Finland between January 1, 1985 and December 31, 2005. There are six variables in this dataset, and table 1 below provides brief descriptions of each variable.

Variable	Description
id	Participant dentification Number
sex	Gender. (1: Female, 0: Male)
age	Age at Diagnosis (in years).
stage	TNM stage of tumor. (1: Stage I; 2: Stage II; 3: Stage III; 4: Stage IV; unkn: Unknown)
time	Survival Time (in years).
event	The status of the patient. (1: Alive; 2: Died of OSCC; 3: Died of other reasons)

TABLE 1. Data Description

2.2. **Missing Values.** The only variable that contains missing values is 'stage', which indicates the TNM stage of the tumor per patient with larger stage numbers representing worse disease conditions. There are 71(21%) patients with unknown stages. Considering that the TNM stage may be an important factor that are related to the death, it is of great importance to explore the cause and consequence of the missingness. A preliminary Kaplan-Meier estimator is constructed to model the survival function. This non-parametric estimator the fraction of patients living for a certain amount of time after joining the study, which will

Date: March 25, 2020.

generate a plot that approaches the true survival function for that population. The following plot (figure 1) demonstrates the Kaplan–Meier curves of the stage cohort, stratified according to data missingness. The figure reveals the survival probability since intake for the group with observed TNM stage records (green) and the group with missing TNM stage records (red). We can observe that the missingness is strongly related to survival because the survival probability of patients without TNM stage records is lower than those who with observed TNM stage records in most time points.

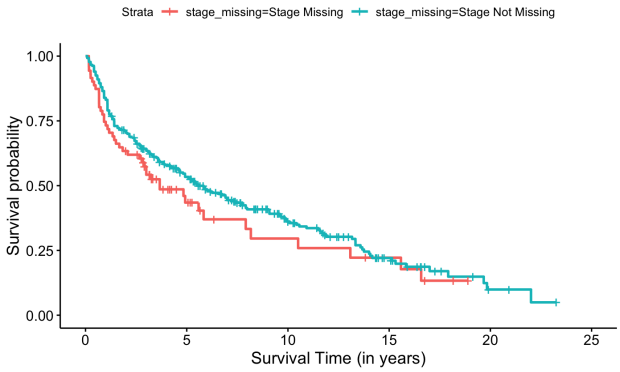

FIGURE 1. Survival Plot for Data Missingness in TNM Stage

Variable	χ^2 **Test**	**Fisher's Exact Test**	**Observed Stage**	**Missing Stage**
Age Group (years)	p = 0.964	p = 0.85		
0-20			0 (0%)	1 (100%)
20-40			5 (23%)	17 (77%)
40-60			23 (20%)	90 (80%)
60-80			32 (21%)	124 (79%)
80+			11 (24%)	35 (76%)
Gender	p = 0.2345	p = 0.2271		
Female			125 (28%)	27 (18%)
Male			142 (86%)	44 (24%)

TABLE 2. Distribution of Demographic Variables over TNM Stage

To investigate if there are specific patterns among patients without observed TNM stage records, tables 2 shows the distributions of variable age and gender over missingness of TNM stages records. Except the one patient who was diagnosed at 15 years old, in the first group, it could be observed that the proportion of missingness among age groups are very close. Furthermore, the large p-values generated from both χ^2 test and Fisher's exact test further confirm that the association between the missingness of TNM stage and age is not significant. When it comes to the gender, the proportion of missingness in male appears to be higher than that in female. However, the p-values from χ^2 test and Fisher's exact test again indicate the insignificant association between TNM stage and gender. Since the proportion of missingness is still lower than the 25% threshold, recklessly dropping subjects without observed TNM stage from the analysis will not be a good decision.[1] From here, the data will be assumed to be missing at random (MAR), which occurs when the missingness is not random, but missingness can be fully accounted for by variables where there is complete

information. In this case, missing values will be imputed using the method of multiple imputation with the R package 'MICE', which generates imputed datasets based on observed values.

2.3. **Exploratory Data Analysis.** Of the 338 patients, there are 185 males (55%) and 152 females (45%).The mean age at diagnosis was 64 years, ranging from 26 to 92 years. From plot [a] of figure 2, we can observed that OSCC is typically diagnosed in their fifth to seventh decade of life ($n = 232; 69\%$). For the patient who was diagnosed at the age of 15, since differences have been identified for the profile of patients with early-onset of OSCC, which suggests that OSCC in the young may be distinct from that occurring in older patients with a different etiology and disease progression, and there only exist one patient below 20 years old, this data point will be treated as an outlier and removed. In terms of patients' status at survival time, 122(36%) patients were directly died of OSCC, with the rest two thirds of patients are either alive ($n = 109; 32\%$), or death of other reasons (n = 107; 32%). Plot[b] of figure 2 explores the relationship between the genders and status of patients, the proportion of alive females is slightly higher than that of men, while the percentage of females who were directly died of OSCC is also greater than that of males. However, overall the percentages of status are close between genders.

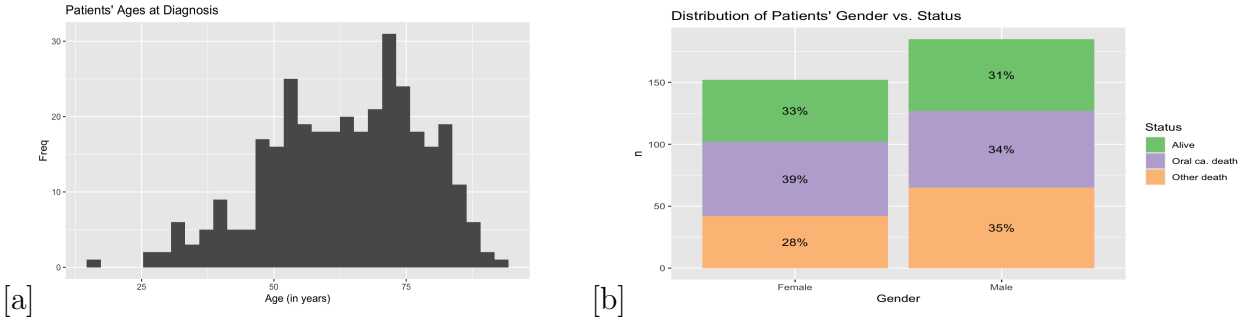

FIGURE 2. Distribution of Age; Distribution of Gender vs. Status

For the distribution of patients' TMN stages, plot [a] in figure 3 reveals that the counts of patients at each stage, with the majority of patients being at stage 2 ($n = 148; 44\%$) and the minority at stage 1 $n = 49; (15\%)$. Furthermore, patients at each TNM stage are classified based on their status at the end of study. Observing from the length of the green part within the stacked bar in the plot, which represents the percentage of patients who were still alive within each stage respectively, we can approximately get the insight that the survival rate sharply decreases as the stage getting larger. In the meantime, specifically from the purple part of bar, it can be observed that more patients directly died of OSCC as stages getting larger as well. If observing the status of patients from a different perspective - survival time, it can be seen from plot [b] in figure 3 that over half of the patients can survive for 5 years (green: dead; red: alive).

3. Non-Parametric Survival Analysis

To evaluate the survival experience of the population across different prognostic factors, a Kaplan-Meier estimator is employed to conduct the analysis, which is a powerful tool for univariable comparison over two/multiple study groups. In this case, patients who passed away from other causes will not be excluded since the focus is the overall survival experience over the whole OSCC population in this study and the data collector did not deny that

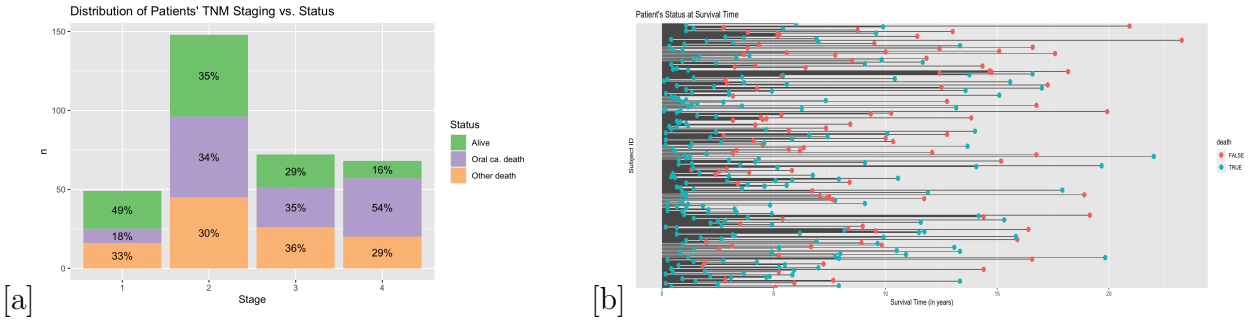

FIGURE 3. Distribution of TNM Stage vs. Status; Display Status at Survival Time

deaths from other causes was related to OSCC.Therefore, the 'event' in this study is defined as death. First of all, a Kaplan-Meier estimator is applied to all patients. Based on the quantile table of the KM fit, there is 75% of chance that an OSCC patients will survive for 1.3 years after diagnosis, and 50% of chance with survival for 5.3 years, which echoes with our finding from plot[b] of figure 3. To explore the study sample in a more thorough way, prognostic factors including gender, TNM stage, and age will be involved in non-parametric survival analysis in the following subsections.

3.1. **Analysis of Gender.** Numerically, the differences between gender proportion among three status groups are not obvious with 46% of women in the alive group and 44% of women in the dead group. To assess whether the survival experience between two genders is significantly different, a Kaplan-Meier estimator is utilized again to compare survival experiences of two genders. Examining the quantiles from this model, it appears that there is 50% of chance that a man will survive for at least 6.9 years, while this reduces to 4.7 years for a woman. A survival plot for gender is provided (figure 4) to illustrate the survival function, which is the probability of surviving beyond any given time points. As a complement, a cumulative hazard plot in also provided, which visualizes the cumulative hazard $H(t_i)$ versus the time t_i of the $i-th$ death.[2] It can be observed that the confidence intervals of two groups overlapped a lot, except that the discrepancy is relatively larger at the period between the second and eleventh years and after 20 years.

To consolidate the phenomenon, a log-rank test is implemented, which is a non-parametric test to test the null hypothesis of no difference in survival between two or more independent groups. It compares the entire survival experience between groups and can be utilized as a test of whether the survival curves are identical (overlapping) or not.[4] The χ^2 statistic is 0.9 with 1 degree of freedom and the test returns a p-value of 0.3. Selecting $\alpha = 5\%$ as the significance level, it therefore confirms that the difference in the survival experience between two genders is not significant. Men generally are not expected to survive significantly longer than women.

3.2. **Analysis of TNM Stages.** Apart from gender, the TNM stage is an important prognostic factor in cancer survival research, which helps to understand how serious the cancer is and the chances of survival. Typically, stages from 1 to 4 indicate that cancer is present and the higher the number, the larger the cancer tumor and the more it has spread into nearby tissues.[3] Following the same procedure as the non-parametric survival analysis of gender, a Kaplan-Meier model is constructed to model the survival function using the imputed data. By checking the model quantiles and summary, we can notice that, in general, patients diagnosed with lower TNM stage tend to have lower mortality rates than patients

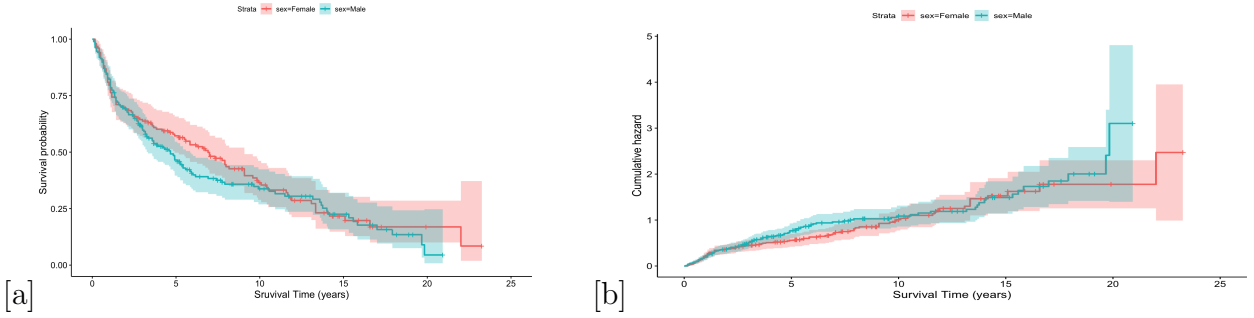

FIGURE 4. Survival Curves of Genders

with higher stage tumors. For example, there is 50% of chance for a patient at stage 1 to survive for 5.6 years, which the chance of that survival years for a patient at stage 5 is less than 25%. Interestingly, the difference between stage 2 and stage 3 is relatively less obvious. Observing the survival curves of TNM stages in figure 5, it is easy to notice that there are overlaps among the first stage, while there exist apparent difference between stage 4 and other stages. To verify whether there exist significant difference among four stages, a long-rank test is implemented and return a p-value $= 0.006$, which is higher than our designated level of significance ($\alpha = 5\%$).

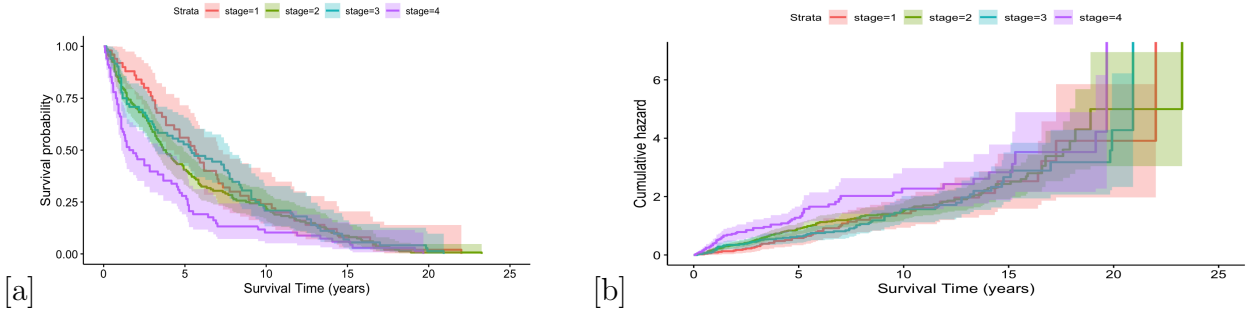

FIGURE 5. Survival Curves of TNM Stages

Reflecting on the result, a potential cause is the phenomenon observed from the KM quantiles that the difference between stage 2 and stage 3 is trivial. After combing stage 1 and stage 2 into a single stage named as "stage 1_2", we refit the KM estimator and acquired the following the survival curves as shown in figure 6. It can be observed that the differences among three groups become obvious. A log-rank test is implemented again and returns a small p-value $= 5 \times 10^{-5}$, which indicates that there exists significant difference among patients in various TNM stages. Since the curves appear to parallel at most time, the assumption of proportional hazards may hold for the Cox PH model in the regression section.

3.3. **Analysis of Age.** Age is also regarded as a important prognostic factor in OSCC clinical research. Before constructing a Kaplan-Meier estimator to model the survival function over age, patients are partitioned into four groups: $20-40$, $40-60$, $60-80$, and $80+$ based on their age at diagnosis. From the quantiles of this KM estimator, patients aging from 40 to 60, who occupy the majority of OSCC population, have 50% of chance to survive for nearly 7 years, and half of those who are aging from 60 to 80 are expected to survive for 5.5 years. However, the chance for people who are over 80 years to survive for more than 5

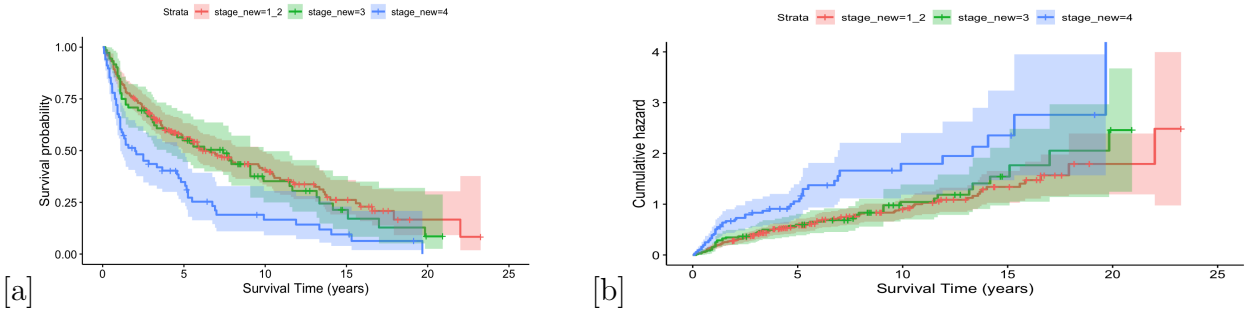

FIGURE 6. Survival Curves of Modified TNM Stages

years is less than 25%. figure 7 demonstrate the survival probability curve and cumulative hazard curve across age groups. It can be observed the difference among each age groups is getting larger over time, and the confidence interval of each age group rarely overlap, which is a signal indicating the difference among age groups may be significant. The p-value ($p = 3 \times 10^{-11}$ returned by a log-rank test confirms the differences among age groups are significant. In the meantime, since the curves do not appear to cross over each other, the assumption of proportional hazards may hold for the Cox PH model in the regression section.

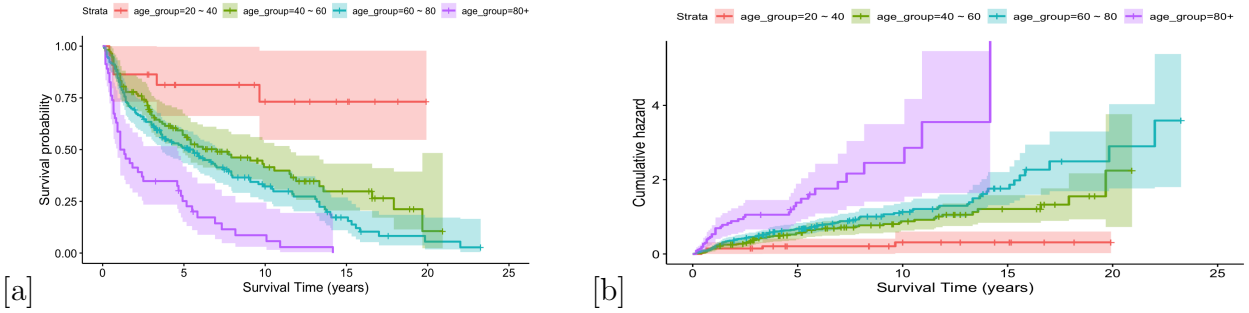

FIGURE 7

4. **Regression Survival Analysis**

Non-parametric tests are particularly feasible when comparing survival functions at the factor level. They are very powerful, efficient and simple. However, the above utilized methods - Kaplan-Meier curves and logrank tests are basically univariate analysis, which describes the survival according to one factor under investigation, but ignores the impact of any others covariates. On the contrary, a regression model is more flexible to explore the relationship between survival and predictors.

A Cox proportional hazards (Cox PH) model will be adopted to explore the effect of several covariates simultaneously, which is a form of linear regression model because it assumes that a single curve is sufficient to to estimate the survival times. Its assumption of proportional hazards make it possible to estimate the effect parameters without dependence on the hazard function as a semi-parametric survival regression model. Another reason that a Cox PH model is preferred in survival analysis is that it has great advantage of flexibility due to the fact that it makes no distribution assumption for the survival data.

Since an important assumption for the Cox PH model is the proportional hazards assumption, the Kaplan-Meier curves of survival functions for groups between two genders severely cross, which indicates that the proportional hazards assumption may not hold. Therefore, gender will not be used as a covariate here. A Cox PH model with covariates age and stage is fitted to the data. It is important to ensure the validity of our Cox PH model before interpreting the model. Firstly, the proportional hazards (PH) assumption is further checked based on the scaled Schoenfeld residuals. [5] Plot [a] of figure 8 displays the graphs of the scaled Schoenfeld residuals against the time for each covariate. From the graphical inspection, there is no pattern with time. The assumption of proportional hazards appears to be supported for the covariates age and TNM stage. Secondly, to test the influential observations or outliers (we have removed an outlier), the deviance residuals will be visualized, which is a normalized transform of the martingale residual. From plot[b] of figure 7, we can observe that the pattern looks fairly symmetric around 0, which indicates that none of the observations is terribly influential individually.

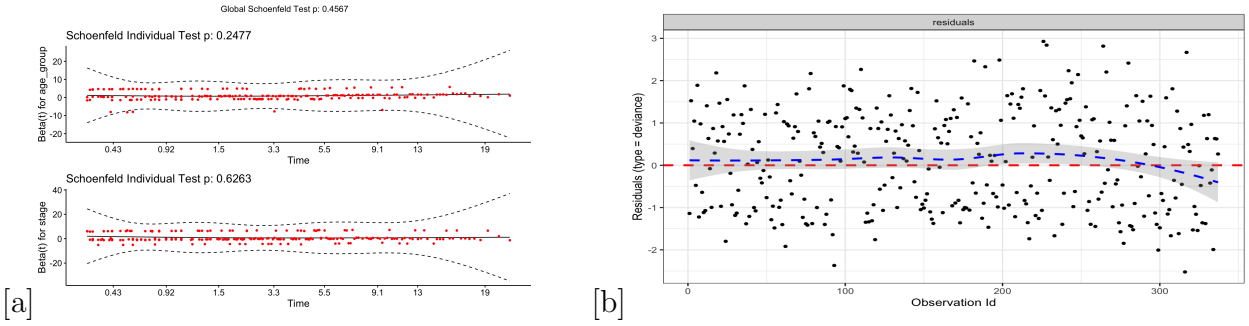

FIGURE 8. Diagnostic Plots of the Cox PH Model

The summary of the Cox PH model is examined after having the model validity confirmed. The extremely small p-values from the likelihood ratio test ($p = 2 \times 10^{-12}$), Wald test ($p = 3 \times 10^{-12}$), and log rank test ($p = 9 \times 10^{-14}$) all indicate that at least one variable in the model has significant effect on the survival times, while further analysis is necessary to determine which variable(s) is (are) crucial. Plot [a] of figure 9 provides an explicit an explicit visualisation for the Cox PH model. Overall, we can find that the effect of age is greater than that of TNM stage with larger regression coefficients. Taking a closer look at each predictor, it can be observed that covariates age and TNM stage are highly significant based on their individual p-values. Furthermore, it is useful to perform variable selection for this model to further investigate the importance of each predictor. Since both predictors (age and TNM stage) are categorical, single term deletion will be implemented, the idea of which is to try fitting all models that differ from the current model by dropping a single term and maintaining marginality. The result of single term deletion shows that deletion of either age and TNM stage leads to great increase in AIC values, which again confirms that both predictors are important.

The Cox PH model can be used to make risk predictions for various survival times with four levels of age ($20-40$, $40-60$, $60-80$, and $80+$) and and three levels of modified TNM stage (stage 1_2, 3, and 4). The ability to predict the risk of an event (death in our case) is great practical use to clinical studies.The constructed Cox PH model to simulated new data was fitted to new simulated data. The twelve curves presented in plot [b] of figure 9 represents the predicted survival rates. We can observe that no matter what TNM stage

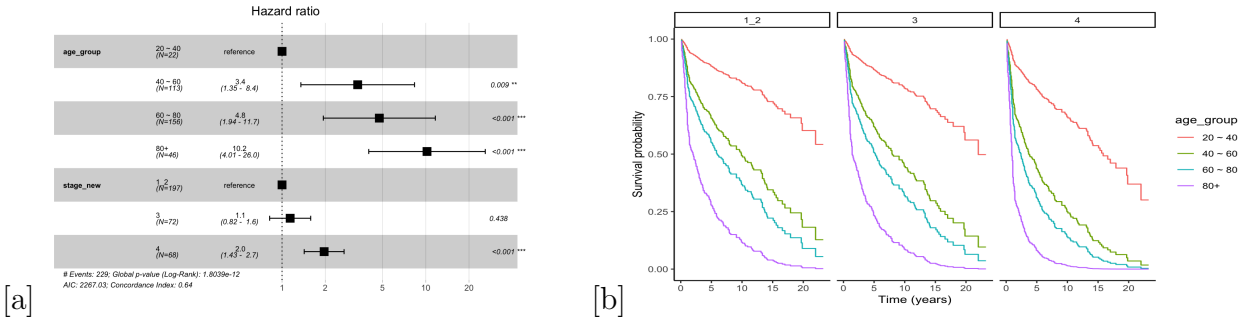

FIGURE 9. Forest Plot for Cox PH Model; Predicted Survival Rates

the patient is in, younger patients always have the highest survival probability in any given time points. The difference between patients aged 40 60 and 60 80 is smaller if compared to difference among other age intervals. In addition, as the number of TNM stages gets larger, the discrepancy among each age groups become smaller. Especially in the stage 4, the survival probability of patients over 40 years old almost converges after 20 years.

The following table (table 3) displays the predicted risk of death across age and TNM stage for the first two years after diagnosis numerically, which is expected to provide some preliminary information for patients who look forward to learn what will happen in the first two year after they are diagnosed. Observing from the table, the risk of death sharply increases from the year 1 to year 2 in all segments. Interestingly, the increase in risk of death across ages are more apparent than that across TNM stages, which echoes with the previous conclusion that age has larger effect than the TNM stage on the survival rate. To have a more straightforward view, the mean and standard error of predicted at each levels of both prognostic factors are calculated in table 4. By comparing the means, we can clear observe the same phenomenon as above.

Age	TNM Stage	Risk of Death at Year 1	Risk of Death at Year 2
20∼40	1_2	3.7%	6.7%
40∼60	1_2	11.8%	20.5%
60∼80	1_2	16.3%	27.7%
80+	1_2	31.8%	50.2%
20∼40	2	4.2%	7.5%
40∼60	2	13.4%	23.0%
60∼80	2	18.4%	30.9%
80+	2	35.3%	54.8%
20∼40	3	7.1%	12.6%
40∼60	3	21.9%	36.3%
60∼80	3	29.5%	47.2%
80+	4	52.8%	74.6%

TABLE 3. Predicted Risk of Death in the First Two Years

	Year 1		Year 2	
TNM Stage	mean	sd	mean	sd
1_2	15.9%	11.8%	26.3%	18.2%
3	17.8%	13.1%	29.1%	19.7%
4	27.8%	19.1%	42.7%	25.8%
Age				
20∼40	5.0%	1.8%	8.9%	3.2%
40∼60	15.7%	5.4%	26.7%	8.5%
60∼80	21.4%	7.1%	35.3%	10.4%
80+	40.0%	11.3%	59.9%	12.9%

TABLE 4. Descriptive Statistics for Predicted Risk

It is also important to assess the model's "goodness of fit" and then one can get to know the prediction power of the regression model. A commonly-used means of checking the fit of a model is the Cox-Snell residual plot.[6] The idea of this diagnostic is based on fitting a Kaplan-Meier (or Nelson-Aalen) curve to the Cox-Snell residual and comparing it to that of the standard exponential. From figure 10, it can be observed that apart from a few values on the left top, most points fall approximately lie on the diagonal, which suggests that the constructed Cox PH model is adequate.

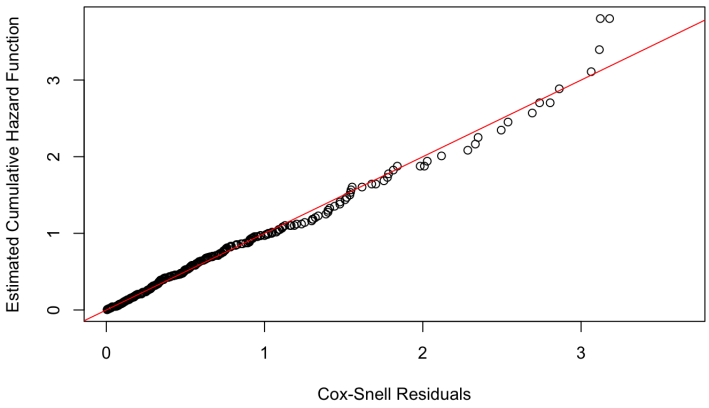

FIGURE 10. Cox-Snell Residual Plot

5. **Conclusion and Discussion**

Oral squamous cell carcinoma (OSCC) is the most common type of oral cancer [7]. Prognostic factors that potentially affect the survival rates was examined in this study. The non-parametric survival analysis establishes that statistically significant association with survival rates are found in age and TNM, while gender does not demonstrate statistically significant affect on effect, which echoes with the preliminary result in the exploratory data analysis. By interpreting the result of the constructed Cox PH model, age has the most significant effect on patients with oral cancers with higher hazard ratio for overall survival. Specifically, having younger patients at 20 40 years old as the baseline, higher hazard ratios are earned by elders patients at 40 60 (3.3 times, $p = 0.009$), patients at 60 80 (4.8 times, $p = 6.63 \times 10^{-4}$), and patients over 80 years old (4.8 times, $p = 1.09 \times 10^{-6}$). In addition, strong

association between the TNM stage and survival rates is also found. Having the combined TNM stage 1 and 2 as the baseline, stage 4 comes with statistically significant high hazard ratio (2.0 time, $p = 3.50 \times 10^{-5}$, while stage 3 does not demonstrate statistically significant effect on the survival rates with a large p-value (0.43). It would be biased to judge the significance of a predictor if one only pay attention to the p-value of individual levels within a predictor.

In summary, a higher TNM stage of the cancer, increased age of the patient at presentation showed to alter the survival grossly. However, I would like to caution readers that it is incorrect to treat it as a precise conclusion because many important factors like treatment, tobacco and alcohol use, and other traditional prognostic factors are not provided in the dataset. Furthermore, despite the flexibility and widely use of the semi-parametric Cox PH mode, parametric regression models may also be efficient if the parametric distributional assumption holds. Some potential candidates of parametric survival regression models may include the accelerated failure time model (AFT) and Weibull survivial model.

References

[1] Buuren, Stef Van. *Sensitivity Analysis.* Flexible Imputation of Missing Data. Second Ediction. https://stefvanbuuren.name/fimd/.

[2] NIST/SEMATECH. *Hazard and cumulative hazard plotting.* e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/.

[3] National Cancer Institute. *Cancer Staging.* https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

[4] LaMorte, Wayne W. *Comparing Survival Curves.* Survival Analysis. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704$_s urvival/BS704_S urvival5.html$.

[5] STHDA. *Cox Model Assumptions.* Statistical Tools for High-throughput Data Analysis. http://www.sthda.com/english/wiki/cox-model-assumptions.

[6] Xu, Ronghui. *Assessing the Fit of the Cox Model.* University of California, San Diego. https://www.math.ucsd.edu/ rxu/math284/slect9.pdf.

[7] Pires, Fábio Ramôa et al. *Oral squamous cell carcinoma: clinicopathological features from 346 cases from a single Oral Pathology service during an 8-year period.* Journal of applied oral science : revista FOB, 21(5), 460–467. https://doi.org/10.1590/1679-775720130317.

Department of Statistics, University of British Columbia