Summer 8-15-2016

Survival Analysis in A Clinical Setting

Yunzhao Liu
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics
Statistics

Survival Analysis in A Clinical Setting

Arts & Sciences Graduate Students
by
Yunzhao Liu (Catherine)

A thesis presented to the
Graduate School of Arts & Sciences
of Washington University in St. Louis
partial fulfillment of the
requirements for the degree
of Master in Arts

August 2016
St. Louis, Missouri

Table of Contents

List of Tables……..iv
List of Figures…….....vi
List of Equations……....vii
Acknowledgement……..ix
Abstract…….....xi
Chapter 1: Introduction……1
1.1 Head and Neck Squamous Cell Carcinoma (HNSCC)……..1
1.1.1 Overview……1
1.1.2 Characteristics……1
1.1.3 Treatment……2
1.2 Human Papilloma Virus (HPV)…….3
1.2.1 Overview……3
1.2.2 Disease Statistics……4
1.3 Risk of HNSCC and HPV……...4
1.3.1 Overview……....….4
1.3.2 Influential Factors……...4
1.4 Survival Analysis……5
1.4.1 General View of Survival Analysis……5
1.4.2 Survival Models……..6
Chapter 2: Methods and Procedures……..17
2.1 Purpose……17
2.2 Dataset……..17
2.2.1 Overview……...17
2.2.2 Categorical Variables……17
2.2.3 Numerical Variables…….……....….19
2.3 Software……...19
2.4 Procedures……20
Chapter 3: Results……..22
3.1 Cox Proportional Hazard Model……..……22

3.1.1 Whole Model……22
3.1.2 Backward Elimination……...23
3.1.3 Forward Selection…….....25
3.1.4 Class P16……...27
3.2 Kaplan Meier Model……28
3.2.1 Whole Model…….28
3.2.2 Strata P16…….……..……...30
3.2.3 Other Factors…….……35
3.3 Accelerated Failure Time Model……..…....43
3.3.1 Whole Model with Various Distributions…….....43
3.3.2 Fit Statistics Comparison…….....44
Chapter 4: Discussion……45
References…….…….51

List of Tables

Table 1……4

Table 2……10

Table 3……10

Table 4……14

Table 5……18

Table 6……19

Table 7……22

Table 8……22

Table 9……23

Table 10……23

Table 11……24

Table 12……24

Table 13……25

Table 14……25

Table 15……26

Table 16……26

Table 17……27

Table 18……27

Table 19……27

Table 20……28

Table 21……29

Table 22……30

Table 23……31

Table 24……31
Table 25……32
Table 26……36
Table 27……37
Table 28……37
Table 29……38
Table 30……40
Table 31……42
Table 32……43
Table 33……44
Table 34……44

List of Figures

Figure 1……..……11

Figure 2……..…12

Figure 3……..…14

Figure 4……..…28

Figure 5……..…30

Figure 6……..…33

Figure 7……..33

Figure 8……..…34

Figure 9……..34

Figure 10……35

Figure 11……35

Figure 12……36

Figure 13……37

Figure 14……38

Figure 15……39

Figure 16……40

Figure 17……41

Figure 18……42

Figure 19……43

List of Equations

Equation 1……6

Equation 2……6

Equation 3……7

Equation 4……7

Equation 5……7

Equation 6……7

Equation 7……8

Equation 8……9

Equation 9……9

Equation 10……..9

Equation 11……10

Equation 12……10

Equation 13……11

Equation 14……11

Equation 15……11

Equation 16……12

Equation 17……13

Equation 18……13

Equation 19……13

Equation 20……14

Equation 21……14

Equation 22……14

Equation 23……15

Equation 24……15
Equation 25……15
Equation 26……16
Equation 27……44
Equation 28……44

Acknowledgments

First and foremost, I thank my thesis advisor Dr. Spitznagel, without his patience and guidance; this thesis would not be possible.

Secondly, I greatly appreciate the acceptance and understanding of our PI, Dr. Virgin, my manager and all my coworkers. Their understanding allowed me to work fulltime while acquiring this degree.

Third, I would like to thank all the hard work that my teachers spent on me. Without you, I would not be able to accelerate.

Last but not least, I want to thank my parents for their unconditional love and never giving up on me. I appreciate all these years that you helped me to become a better person.

Yunzhao Liu (Catherine)

Washington University in St. Louis

August 2016

Dedicated to my late grandparents whom gave me the inspiration to make a positive impact on society; love you with all my heart.

Abstract

Application of survival Analysis in A Clinical Setting

for Arts & Sciences Graduate Students

by

Yunzhao Liu (Catherine)

Master of Art in Statistics

Mathematics

Washington University in St. Louis, 2016

Professor Edward Spitznagel, Chair

Professor Todd Kuffner, Co-Chair

Professor Guoyan Zhan, Co-Chair

With the fast paced advancement of modern medicine, cancer treatments have improved greatly over the past few decades; however, the overall survival rate has not improved for head neck squamous cell carcinoma (HNSCC). Traditionally, the general affected population of HNSCC was male over 50-60 years of age, whom have had history of alcohol and tobacco use. Conversely, in the recent decades, HNSCC has exhibited significant rise in younger patients, largely due to the increase in human papillomavirus (HPV) infection among young adults. Generally, HPV as the most prevalent sexually transmitted disease, consisted of strains that do not cause harm to humans. Only handful of strains were found to be carcinogenic, potentially. Furthermore, the carcinogenic property of HPV has been increasing tremendously, and becoming a greater threat to human. For instance, HPV is the leading cause of cervical cancer currently. Recently, HPV related HNSCC has showed significant increase in the last 30 years as well, with

oropharyngeal squamous cell carcinoma (OPSCC) as the most prevalent type, and the most increased kind in the HPV related HNSCC groups.

In this study, three methods of survival analysis were used which included non-parametric Kaplan-Meier method, parametric accelerated failure model and Cox proportional hazard method to achieve this data analysis.

First, two best fitted predictive survival models were developed for HNSCC (OPSCC) patients whom have been diagnosed and treated at Barnes Jewish Hospital in St. Louis. The models were initially determined by forward and backward selection of Cox proportional hazard method. The best predictive variables were further identified via forward selection in Kaplan Meier method. As a result, the final model estimates were obtained through accelerated failure time model.

Additionally, using Kaplan Meier method, HPV and HNSCC (OPSCC) relationships were investigated via P16 protein presence, which is an indicator of HPV related OPSCC. Survival rate of P16+ and P16− status were compared and contrasted. Interaction between the presence of P16 protein and other factors such as age groups, tobacco use, loco-regional fail, various stages of cancer defined by tumor differentiation, cancer recurrence, and lymph node found positive for cancer were explored.

Lastly, other factors of interest such as types of treatment, types of chemotherapy, race and anemia were investigated for overall survival rate as well as interactions with presence or absence of P16, also using Kaplan Meier method. Survival graphs were generated for the whole model as well as for the group comparisons.

Chapter 1: Introduction

1.1 Head and Neck Squamous Cell Carcinoma (HNSCC)

1.1.1 Overview

As the seventh most common cancer, approximately affecting 600,000 people worldwide and accounts for 3% of all cancers, head and neck squamous cell carcinoma (HNSCC) is defined as cancers which affect squamous cells in the mucosa membranes around the nose, mouth and throat region. More specifically, the regions include the oral cavity, oropharynx, nasopharynx, larynx, and hypopharynx [18].

HNSCC affects male around 50-60 years old historically. However, recently cancer cases of younger people are on the rise. Around 75% of HNSCC are the result of tobacco and alcohol use, which mostly are within the older group [10,11]. Recently, Human Papillomavirus (HPV) is becoming a significant factor that can increase the chance of developing HNSCC.

1.1.2 Characteristics

Depending on the causations of HNSCC, this cancer utilizes different carcinogenic pathways. HNSCC associated with tobacco and alcohol use is characterized by P53 mutation, and more prevalent in older patients over the age of 50-60 years old. While HNSCC associated with HPV is characterized by P16 mutation, which resulted in the increase of P16 protein expressions. A protein called E7 in HPV causes pRb degradations, which leads to the overexpression of P16 protein in the host [4, 5, 23].

Several indications or significant factors are related to HNSCC. For none-HPV related HNSCC group, characteristics include anemia, tobacco and alcohol use, ACE27 index, and race have been found significant in this subgroup [1, 2, 21]. Anemia is characterized by reduced red blood

cell count, hemoglobin (Hgb), and hematocrit, which is another way to measure red blood cell count. It has been previously found to be prevalent within the HNSCC population, and suspected to be related to the presence of cancer or comorbid diseases. Also anemia is traditionally correlated to smoking which is the cause for P53 related (non-HPV related) cancer [1, 2]. One way to access comorbidity mentioned above is the Adult Comorbidity Evaluation-27 (ACE27). ACE27 is an index that ranks the severity of comorbidity, which is defined as the presence of 2 diseases simultaneously. Study has found that P16− patients exhibit more comorbid diseases than P16+ group. In the study, 43.3% of P16− patients had severe disease compared to a much less percentage of P16+ patients. Comorbidity was also found to be more prevalent in current smokers in the same study [8,13].

1.1.3 Treatment

Several treatments are available for HNSCC. First and the most prevalent treatment is primary surgery to remove the tumor, others include chemotherapy (CT), chemo-radiation therapy (CRT) and radiation therapy (RT). Initially, the typical treatment suggested by physicians is primary surgery, unless the cancer tumor is miniscule, in which case, CT, RT, or CRT is recommended. Following primary surgery, CT, RT or CRT is often suggested as follow-up treatment. Furthermore, Chemotherapy treatment consists of a group of drugs which target cancer cells. Within Chemotherapy, there are induction chemotherapy and concurrent chemotherapy for this current study [17].

Moreover, the goal of RT is to deliver a lethal dose of radiation to the target tissue and consequential surroundings. Several radiation therapies were conducted for the study which included various types of intensity modulated radiation therapy (IMRT), external beam treatment

involving emission of photon or electron, or combination of both, and definitive radiation treatment [15].

Lastly, CRT is the combination of CT and RT, which is found to be effective for HNSCC (OPSCC). Many times, doctors would offer a combination of above treatments to optimize patient's chance at survival [15, 16].

1.2 Human Papillomavirus (HPV)

1.2.1 Overview

Human Papillomavirus (HPV) comprises of a group of DNA viruses which have the potential to infect basal epithelial cells, both skin and mucosal layer. HPV consists of ~200 strains, and estimated to be the most prevalent sexually transmitted disease. Certain strains of HPV are able to trigger genital warts, and various types of cancers such as cervical cancer, penile cancer, oropharyngeal cancer and others. However, only a small percentage of the strains are associated with genital warts and cancers. More specifically, about 40 strains can infect the genital, mouth, and throat area in men and women. furthermore, the strains that is responsible for genital warts are different from the ones that cause cancer [7].

1.2.2 Disease Statistics

How prevalent is HPV? According to CDC, about 79 million Americans are infected with HPV, and 14 million are infected each year. HPV has been the leading cause of cervical cancer in women, and it is predicted to affect approximately 500 thousand women worldwide [18, 23]. The carcinogenic property of HPV is increasingly becoming a greater risk for HNSCC and described successively.

1.3 HNSCC Risk and HPV

1.3.1 Overview

HPV related HNSCC is at a steadily incline for the pass 30 years, which contributed to increase cancer risk of young individuals with HPV infection, especially male. HPV is detected in about ¼ of all HNSCC, with majority of them being oropharyngeal squamous cell cancer (OPSCC) which is one of the most rapid growing cancer currently [6, 9].

1.3.2 Influential Factors

Recent studies have shown that HPV is associated with various types of head and neck squamous cell carcinoma (HNSCC), with OPSCC as the most increased and consists of majority of the HPV related HNSCC group. Furthermore, cancer risk is more prevalent in developing country compared with developed countries [7]. In table 1, different types of HPV related cancer and the relating statistics are presented below:

Site	Attributable to HPV (%)	Developed countries		Developing countries	
		Total cancers	Attributable to HPV	Total cancers	Attributable to HPV
Cervix	100	83 400	83 400	409 400	409 400
Penis	40	5 200	2 100	21 100	8 400
Vulva, vagina	40	18 300	7 300	21 700	8 700
Anus	90	14 500	13 100	15 900	14 300
Mouth	> = 3	91 200	2 700	183 100	5 500
Oro-pharynx	> = 12	24 400	2 900	27 700	3 300
All cancers	5	5 016 100	111 500	5 827 500	449 600

Table 1 Adapted from Parkin et al. 2002

The table above consists of cancer statistics up to 2002, which is concurrent with the time frame which this present study was conducted. However, HPV related HNSCC (OPSCC) is much higher by 2016.

HPV can be identified via the overexpression of P16 protein in HNSCC (OPSCC), as mentioned in 1.1.2. P16 has established as surrogate marker for HPV+/ OPSCC patients. However, the identification is not limited to P16 prevalence. In a study of 496 patients done by Robinson et al.

have found that only 5% were HPV−/P16+, and 8% were HPV+/P16−. P16− negative patients were significantly more frequently anemic than p16+positive patients [2, 24].

Additionally, studies have shown that P16+ patients have better survival rate than P16− patients for HNSCC (OPSCC) subgroup. Also, P16+ patients are usually younger with better socio-economic status than P16−; since P16+ is associated with HPV related cancer rather than alcohol and tobacco related which can have an impact on socio-economic status [5, 9, 22].

Difference in race has also been found amongst HNSCC patients. For HPV+/P16+ group, Caucasian (67%) was found to be more prevalent then African American (25%) patients. Other study has found that HNSCC has worse mortality rate for African American Patients compared with Caucasian patients [3].

1.4 Survival Analysis

1.4.1 General View of Survival Analysis

One question arises regarding Survival Analysis is why should one choose this form of analysis versus ordinary least squared and/or other regression methods. The answer lies within the inability of ordinary regression models at handling censored or truncated data. Conversely, survival analysis has the capability to handle the influence of time, and censored or truncated data. Survival analysis is designed to investigate time at which an event occurs (event time). The events typically involve death of an individual, incidence of certain disease, failure of machinery and other similar natured occurrences.

In survival analysis, three common types of censoring are often discussed, which are left, right and interval censoring. Right censoring is when an observation is dismissed before the event happens. Left censoring is when the event of interest has happened before the data is collected. Interval censoring is when an observation has happened during the time of the study, however

without knowing the exact time, thus lost the ability to be present in the dataset. Right censoring occurs more frequently than left censoring in survival analysis.

One attribute of survival analysis is the ability to calculate hazard, and is essential to survival analysis. The hazard function is shown below:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\left(t \le T < t + \Delta t \mid T \ge t\right)}{\Delta t} \qquad \text{Equation 1}$$

The goal of hazard function is to calculate instantaneous risk that an event will happen at time t. From the hazard function, the survival function can be formulated. The survival function calculates the probability of an individual surviving beyond a given time t. A simple form of the survival function is presented by:

$$S(t) = exp\left[-\int_0^t h(u)du\right] \qquad \text{Equation 2}$$

Furthermore, three methods are most popular amongst survival analysis, which are Kaplan Meier method, accelerated failure time model, and Cox proportional hazard method. They are described subsequently[12, 19].

1.4.2 Survival Models

Regression Kaplan-Meier Method

Kaplan Meier method is a non-parametric, one sample method, which does not assume a distribution. It measures survival probability over time, without making assumption of proportionality.

In Kaplan Meier method, the Kaplan Meier (KM) estimator is a widely used tool, especially in biomedicine. This method is the default function of Proc Lifetest in SAS [12, 14, 19]. KM estimator is a nonparametric maximum likelihood estimator, also known as the product limit estimator. KM estimator is defined as:

$$\hat{s}(t) = \prod_{j:t;\leq t}\left(1 - \frac{d_j}{n_j}\right)$$ Such that: $t_1 \leq t \leq t_k$ Equation 3

The equation presents that at any given time t, the estimator is all the events that occurred during the elapsed time from 1to j. This basically is a survival estimate of the conditional probability of starting time to end time t_{j+1}. Another way to look at the equation is:

$$\hat{s}(t) = \begin{cases} 1, & t > tk \\ \prod_{j:tj\leq t}\left(1 - \frac{d_j}{n_j}\right), & t \leq tk \end{cases}$$ Equation 4

This means that when $t>t_k$, the result is 1; otherwise, the equation can be estimated via the KM estimator.

Another advantage of the Kaplan Meier method is the ability to test over various strata. When strata are being examined, the KM estimator separates the result table by each stratum, and survival graphs provide a curve for each stratum for comparison, which is mentioned later in the section. Within stratified Kaplan Meier method, 3 tests are available for the hypothesis testing, and illustrated in the subsequent analysis. The tests are log rank, Wilcoxon, and -2 log (LR).

Log Rank test is the most widely used and the equation is defined as:

$$\sum_{j=1}^{r}(d_{1j} - e_{1j})$$ Equation 5

This equation presents that the sum of all the event times in all strata over total time r.

Wilcoxon test only differ from the log rank test by multiplying by n (sample number) and given by:

$$\sum_{j=1}^{r} n_j(d_{1j} - e_{1j})$$ Equation 6

This implies that the Wilcoxon test is a weighted test, which results in giving the earlier event more weight compared with later events. This test is more powerful when the event time possess a log-normal distribution.

Lastly there is the −2log (LR) test. This test can be biased because it assumes that the hazard function is constant in every group, and has an exponential distribution [12, 19].

Despite the fact that it's only one sample, Kaplan Meier has many advantages, for example, the ability to generate survival graphs. The two graphs available are the product limit survival graph and the negative log-log survival graph. The product limit survival graph is a step like graph that shows survival probability at a give time t. The latter is just a simple negative log-log transformation ($log\left(-log\hat{S}(t)\right)$ to the survival probability, and a log transformation for time. This transformation makes the step-like product limit graph more interpretable when graphed with strata. Both graphs are great at illustrating models with strata. The differences between strata can be seen and interpret visibly.

Accelerated Failure Time (AFT) Model

The accelerated failure time model (AFT) is a parametric model, which has the underline assumption that the model follows some known distribution, such as binomial, Poisson or normal distribution. The advantage of assuming a distribution is the ability to see the shape of the hazard functions, which can make subsequent inferences easier to obtain. Another benefit of AFT model is that it can accommodate left and interval censoring while Cox's proportional hazard model which is mentioned in the next section can only handle right censoring.

In SAS, the AFT model is built within Proc Lifereg and all the models within are calculated based on maximum likelihood method. The specific maximum likelihood method that Proc Lifereg uses is the Newton Raphson algorithm which is defined as:

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \mathbf{I}^{-1}(\boldsymbol{\beta}_j)\mathbf{U}(\boldsymbol{\beta}_j) \qquad \text{Equation 7}$$

This algorithm estimates the covariance matrix of the coefficients.

Proc Lifereg uses ordinary least squared (OLS) method to calculate this algorithm and treats the censored data as uncensored.

During hypothesis testing, Proc Lifereg employs a chi-squared test, more specifically the Wald test and the equation is described as:

$$\frac{(\hat{\beta}_3 - \hat{\beta}_4)^2}{Var(\hat{\beta}_3) + Var(\hat{\beta}_4) - 2Cov(\hat{\beta}_3, \hat{\beta}_4)} \quad \text{Equation 8}$$

Wald test examinations wether the coefficients of the corresponding varibles equal to 0 or otherwise. Additionally, Proc Lifereg provides a Lagrange multiplier chi-squared statsitics or simply a score statistic to test if the scale parameter is 1.

Additionally, AFT model has the ability to produce predicted event time for any indicated set of covariate values which lacks in the other models. The AFT model satisfies parameters such that:

$S_j(t) = S_i(\phi_{ij} t)$ for all t (time) Equation 9

This equation implies that the difference between 2 individuals or events is the rate at which they progress over time. For example, for human, it would be the rate they age. S_j is the survival probability of the expected, while S_i is the survival probability of observed, and ϕ_{ij} is a constant describing the relationship.

Furthermore, If the dataset does not have censoring, AFT model estimates variables much like an ordinary linear regression and presented as:

$$logT_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \sigma\varepsilon_i \quad \text{Equation 10}$$

The error term in linear regression is typically assumed to have a normal distribution and since this is a logged equation, the error here has a log-normal distribution. However, many survival datasets have censoring, and AFT model have different distribution to accomendate error term for such senerios and shown below:

Distribution of Ɛ	Distribution of T
Extreme value with 2 parameter	Weibull
Extreme value with 1 parameter	Exponential
Log-gamma	Gamma
logistic	Log-logistic
normal	Log-normal

Table 2

Typical distributions used in AFT modeling are generalized gamma, Weibull, exponential, log-normal, and log-logistic, which are explained further subsequently.

The Gamma Model:

The Gamma model makes the broadest assumption and is typically known as the generalized Gamma model. All following models are nested within the gamma model. The characteristic of gamma distribution is that it possesses a shape and a scale parameter. A table of shape and scale parameters relationships between other distributions and gamma model:

Shape=1	Weibull
Shape=1 and Scale=1	Exponential
Shape=1	Log-normal

Table 3

The Weibull Model:

The Weibull model makes the second broadest assumption. The survival function presents:

$$S_i(t) = exp\left\{-\left[t_i e^{-\beta x_i}\right]^{\frac{1}{\sigma}}\right\} \quad \text{Equation 11}$$

The Weibull has a monotonic hazard function and is shown as:

$$logh(t) = \alpha logt + \beta_0^* + \beta_1^* x_1 + \cdots + \beta_k^* x_k \quad \text{Equation 12}$$

The relationship to OLS regression model is such that:

$\beta_j^* = \frac{-\beta_j}{\sigma}$ for j=1,…,k and $\alpha = \left(\frac{1}{\sigma}\right) - 1$ when $\beta_j = 0$, and if and only if $\beta_j^* = 0$

Equation 13

In this model, when σ>1, hazard is decreased with time. When variance is between 0.5 and 1, the hazard is increasing at a decreasing rate [book]. When the variance is between 0 and 0.5, the hazard is increasing at an increasing rate. When σ=0.5, the hazard function displays a straight ling starting at the origin. Below is a graph illustrating different σ value after it's transformed into α, and the equation presented above:

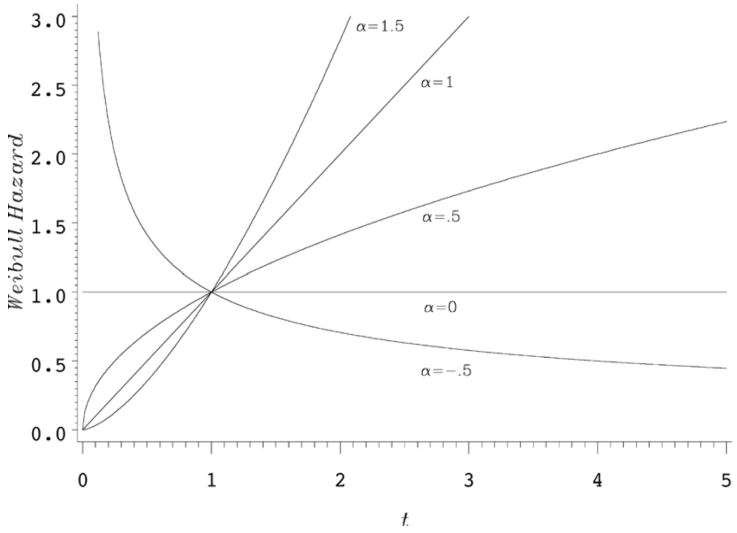


Figure 1 Adapted from Allison et al

The Exponential Model:

As the simplest model within the series, this model assumes constant hazard over time, which is expressed as:

h(t)=λ Equation 14

Expressing the equation in a regression form:

$logh(t) = \beta_0^* + \beta_1^* x_1 + \cdots + \beta_k^* x_k$ Equation 15

When equation 15 is compared with ordinary regression equation (equation 10), $\beta_j = -\beta_j^*$.

There are more assumptions made by this model is that the error has an extreme-valued distribution same as the Weibull model, and contains variance equals to 1, which makes this model is a special case of Weibull distribution. This characteristic will make the scale parameter in Proc Lifereg equal to 1 as seen in table 3. The distribution is not symmetrical and skewed to the left.

The Log Normal Model:

The log normal model has normal distribution with log transformation. It has a non-monotonic hazard function, which is different from the Weibull model. The hazard function is defined as:

$$logh(t) = logh_0\left(te^{-\beta x}\right) - \beta x \qquad \text{Equation 16}$$

This implies that when t=0, the hazard is also 0. Log-normal model is not a proportional hazard model and it does not have a closed form (unscaled normal distribution does not have a closed form). Therefore, l-normal model is presented in a logistic form typically as seen in equation 16. When the variance is large in this model, the hazard peaks rapidly and appears similar to Weibull and Log-logistic models. A graph of different variances with median=1 is presented below.

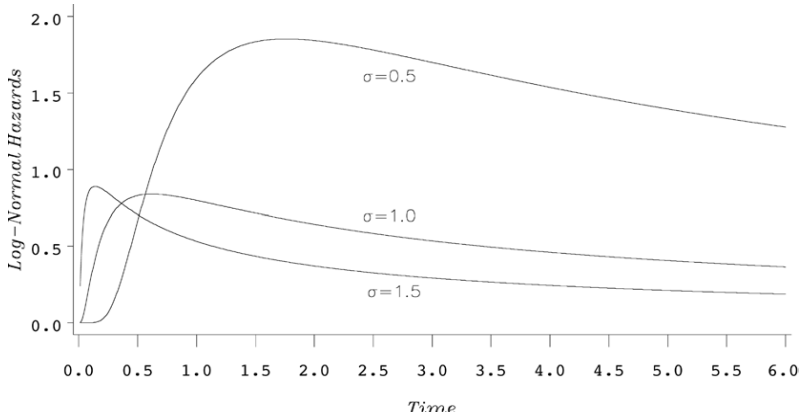


Figure 2 Adapted from Allison et al

This model is best used for repeatable events. For example, in an event of buying a new car, immediately after the purchase, the chance of the same person buying another car is very low. Hence the left screwed peak where the hazard rate initially raises and eventually drops over time.

The log logistic model:

As the name implies, the l-logistic model assumes that its error retains a logistic distribution. Typically, l-logistic model has a dichotomized dependent variable. This model also possesses an inverted U-shaped hazard curve as the l-normal and Weibull model. However, unlike l-normal model, this distribution is symmetrical with a mean of 0.

The log logistic hazard function:

$h(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1+(\lambda t)^{\gamma}}$ where γ=1/σ and $\lambda = exp\{-[\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k]\}$

Equation 17

The Survival Function of L-Logistic Model is seen as:

$S(t) = \frac{1}{1+(\lambda t)}$ with the same restrictions the hazard function.

Equation 18

A logged regression view of the survival function:

$$log\left[\frac{S(t)}{1-S(t)}\right] = \beta_0^* + \beta_1^* x_1 + \cdots + \beta_k^* x_k - \gamma log t$$ Equation 19

$\beta_j^* = \beta_i/\sigma$ for all i=1,…,k compared to ordinary regression. When σ<1, the hazard is similar to the log-normal hazard. When σ>1, the hazard is similar to the decreasing Weibull hazard. When σ=1, the hazard equal to λ at time 0 and eventually declines to 0 as time approaches infinity. A graph of σ over time is presented as:

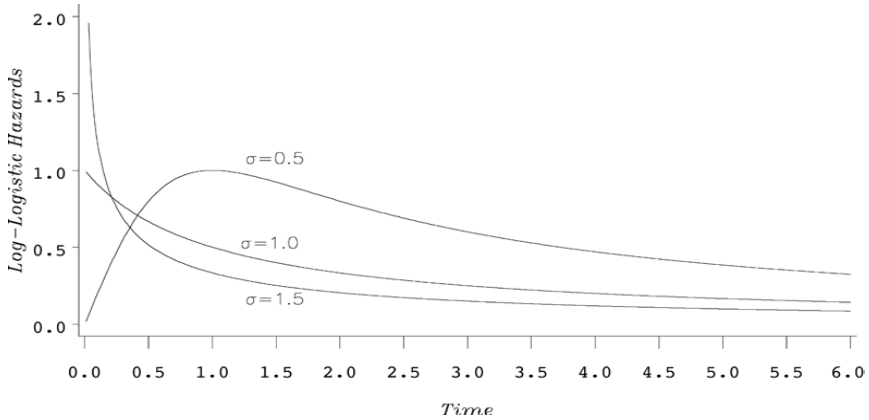


Figure 3 adapted from Allison et al

The log-logistic model is best utilized with binary data such as categorical data of yes or no, and dead or alive, for instances.

Here is a summary table with survival and hazard function of exponential, Weibull and Log-logistic model, seen beneath:

Distribution	$S(t)$	$h(t)$
Exponential	$\exp(-\lambda t)$	λ
Weibull	$\exp(-\lambda t^p)$	$\lambda p t^{p-1}$
Log-logistic	$\frac{1}{1+\lambda t^p}$	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$

Table 4 adapted from Allison et al

Different models are ranked by few fit statistics which are Akaike's Information Criterion (AIC), corrected version of AIC (AICC), and Bayesian Information Criterion (BIC). The equations are seen as:

AIC: AIC=-2logL+2k Equation 20

AICC: AICC=AIC+$\frac{2k(k+1)}{n-k-1}$ Equation 21

BIC: BIC=-2logL+klogn Equation 22

AIC is a modified version of -2 log-likelihood and it penalizes models that have more covariates or more parameters. The benefit of AICC is its adequacy with small samples. Moreover, BIC penalizes for large sample number or additional covariates.

The biggest downfall of Proc Lifereg method is the inability to include time-dependent covariates, in which case, Proc Phreg can be used and described below[12, 19].

Cox Regression-Proportional Hazards Model

Named after Sir David Cox, whom first proposed this method through his paper "Regression Models and Life Tables" in the 1972 issue of Journal of the Royal Statistical Society, Series B, Cox Regression-Proportional Hazard Model has few advantage when compared with the Parametric model presented in Lifereg. This model does not require a distribution as Lifereg, thus making it semi-parametric. Due to this characteristic, Cox's method is more robust than parametric model as well. Moreover, because its semi-parametric property, integration of time-dependent covariates became much easier.

In the software SAS, this model is included in the procedure, Proc Phreg, which has both proportional and non-proportional hazard models. The proportional hazard model is derived from the simple non-proportional hazard model. Below is the equation for non-proportional hazard model:

$h_i(t)=\lambda_0(t)\exp(\beta_1 x_{i1}+\ldots+\beta_k x_{ik})$ Equation 23

Function $h_i(t)$ is the hazard of *i* at any given time t, and represented by a positive baseline hazard function $h_0(t)$ multiplies an exponential of covariates represented by X's, 1 to k.

And a logarithmic version of the same equation presented below:

$\log h_i(t)=\alpha(t)+ \beta_1 x_{i1}+\ldots+\beta_k x_{ik}$ Equation 24

Why it is a proportional hazard? The reason is that the hazard of one person is fixed from hazard of another person and the equation is seen as:

$$\frac{h_i(t)}{h_j(t)} = exp\{\beta_1(x_{i1} - x_j) + \cdots + \beta_k(x_{ik} - x_{jk})\}$$ Equation 25

Proc Phreg utilizes maximum partial likelihood to estimate coefficient β. The benefit of using partial likelihood is that specifying baseline hazard function $h_0(t)$ is no longer needed. Equation of maximum partial likelihood model after maximize β is presented beneath:

$$logPL = \sum_{i=1}^{n} \delta_i \left[\sum_{j=1}^{n} Y_{ij} e^{\beta x_j}\right] \quad \text{Equation 26}$$

The build in method of Proc Phreg which handles tiered models is the Breslow's method. Also, Proc Phreg provides a likelihood ratio test, a score test which is the same as the log rank in Kaplan Meier method, and a Wald test which is discussed in AFT model for tiered data[12, 19]. The benefit of Proc Phreg is the ability to optimize model using backward and forward selection which is utilized in the subsequent analysis. Backward selection considers the full model first, and deletes a predictor with the highest P-value one at a time till the model only consists of predictors that have less or equal to the indicated P-value. Forward selection includes the predictor with the lowest P-value first, then incorporates the next predictor with the lowest P-value until all the predictors that have less or equal to the selected P-values are included in the model [20].

Chapter 2: Materials and Methods

2.1 Purpose

Several aims were established for current study. First, a full predictable survival model with the most suitable predictors is going to be constructed. Second, with the dramatic increase in HPV related OPSCC, P16 status will be investigated by itself, and with interactions of other available predictors. Last, other factors that may provide benefit to patient survival will be investigated as well. The data analysis will not necessarily be in above sequence.

2.2 Dataset

2.2.1 Overview

The dataset consists of information regarding 300 HNSCC (OPSCC) patients obtained from Washington University in St. Louis. All patients in the dataset were treated and diagnosed in Barnes Jewish Hospital in St. Louis, MO, and the follow-ups were done in the same institute. All patients were not previous treated or diagnosed and the data is de-identified for patient privacy purpose. The study was conducted from June 1996 to June 2010, with follow-up through December 2014. Cancer statuses were gained through a database provided by the Department of Pathology, Otolaryngology, and Radiation Oncology. Comorbidity and outcome information were attained from the Oncology Data Services tumor registry. Vital statuses were acquired from electronic medical record which was further confirmed with the Social Security Death Index. There are 19 variables within the dataset and description and variable statistics can be seen in table 5 and 6.

2.2.2 Categorical Variables

Categorical Variables	Descriptions	Categories		%
Anemic	Status of anemia	·	Missing (NA)	10.3
		0	No	74.0
		1	Yes	15.7
Differentiation	Differentiation Stage of cancers	0	None	17.0
		1	Poorly Differentiated	56.7
		2	Moderately Differentiated	24.7
		3	Well differentiated	1.7
Treatment5	Types of treatment received by the cohort	·	Missing (NA)	0.3
		1	Primary conformal radiation therapy	18.3
		2	Only surgery	8.7
		3	Primary radiation therapy	3.3
		4	Surgery+adjuvant radiation therapy	36.0
		5	Surgery+adjuvant conformal radiation therapy	33.3
Radmodality	Modality of radiation therapy	0	None	9.3
		1	Intensity-modulated radiation therapy(IMRT), external beam treatment	34.3
		2	External beam therapy via a photon producing machine with beam energy ranging 6-10mv	12.7
		3	Treatment by external beam	30.0
		4	Definitive radiation therapy	2.0
		5	Post-operative adjuvant radiation therapy	8.3
		6	Post-operative adjuvant IMRT	2.0
		7	COMBINATION SPEC (< 2003)	0.3
		8	Definitive IMRT	0.7
		9	Treatment delivered using a combination of photon and electron beams.	0.3
Chem3	Chemotherapy status of the patients	·	Missing (NA)	0.3
		0	None	48.0
		1	Induction chemotherapy	11.7
		2	Concurrent chemotherapy	40.0
Tobacco3	Tobacco usage	·	Missing (NA)	8.3
		0	None user	23.7
		1	Current user	34.0
		2	Former user	34.0
Alcohol	Alcohol usage	·	Missing (NA)	11.0
		0	No	22.0
		1	Yes	67.0
Recurrence		·	Missing (NA)	8.0

	Recurrence of cancer	0	No	77.7
		1	Yes	14.3
Cancerstatus	Cancer status of the cohort	0	Free of this disease	80.7
		1	Not free of this disease	19.3
Vitalstatus	Status of dead or alive	0	Alive	60.3
		1	Dead	39.7
P16	P16 protein status	0	Negative	23.3
		1	Positive	76.7
Sex	Gender of the cohort	0	Male	87.3
		1	Female	12.7
Race	Race of the cohort	0	Others	1.7
		1	White	86.3
		2	Black	12.0
ACE_27	Adult comorbidity Evaluation 27 index value	·	Missing (NA)	1.7
		0	None	39.7
		1	Mild	37.7
		2	Moderate	14.0
		3	Severe	7.0
Locoreg_fail	Locro-regional failure	·	Missing (NA)	0.7
		0	No	94.0
		1	Yes	5.3

Table 5

2.2.3 Numerical Variables

Variable	Description	Range		Mean	Median	NA's	Units
		Lower	upper				
HGB	Hgb level	8.9	17.2	14.29	14.3	31	gm/dl
Hematocrit	Hematocrit level	8.0	50.0	41.79	41.80	31	RBC%
Durationmo	Time since diagnose	2.8	212.6	71.24	68.85	0	Months
Age	Age of the patient	32.5	87.1	56.27	56.00	0	Years
LN_positive	Lymph nodes tested cancer positive	0.0	40.0	3.25	2.0	72	Counts

Table 6

2.3 Software

Statistical software R was used for variable transformation and data subset. Command ifelse from package {base} was used to transform data into binary or categorical variables. Next

command cbind from the same package was used to combine desirable variables in to working dataset. Subsequently, dataset was exported as comma separated (csv) text via command write.csv in R package {utils}.

SAS 9.4 statistical package was used for all the analysis and modeling. Proc Lifetest was utilized for Kaplan-Meier method. Proc Lifereg was used for AFT method, and Proc Phreg was applied for proportional hazards model.

2.4 Procedures

Initially, the dataset was explored as a whole, where a complete model including all the variables was developed. Cox proportional hazard regression with command Proc Phreg was utilized to achieve in building the full model. The response variable is Durationmo. The censored variable is Vitalstatus=0, and the predict variables were P16, Radmodality, Cancerstatus, Treatment5, LN_positive, Age, Sex, Anemic, Hematocrit, HGB, Chem3, Tobacco3, Alcohol, Recurrence, Locoreg_fail, ACE_27, Differentiation and Race.

Furthermore, backward and forward selections were used to optimize the complete model, and only the significant variables specified at $p \leq 0.15$ for backward selections and $p \leq 0.20$ for forward selections were kept. The commands for those selections were slstay=0.15 and slentry=0.20, respectively. Additionally, Cox proportional hazard regression in Proc Phreg was used for backward and forward selection to determine which factors were significant for P16+ and P16− status separately. Also, P16 was treated as class (class P16) using Proc Phreg for backward elimination to see which factors are significant when P16 was treated as categorical variable and to see if there were any interactive terms for P16.

Next, using Kaplan Meier method with syntax Proc Lifetest, survival graphs were generated for HNSCC survival rate with command Durationmo*Vitalstatus(0). Additionally, whole model

backward and forward selection was validated with test statement in Proc Lifetest (test<variables>), which provided a summary table of parameter estimates of all the variables, and a table of forward selection of each variable. Subsequently, P16 status was investigated as strata over Durationmo*Vitalstatus(0). This was achieved by using command strata P16. Survival probability graphs were also made for each variable. Test statement was used for specifying strata P16 as well. Additionally, P16 was paired as strata with other variables. The variables were Age, Tobacco3, Recurrence, Locoreg_fail, Differentiation, and LN_positive. Moreover, variables that were found significant in the other models in this study, and in the previous studies mentioned in the introduction, or can potentially possess importance in patient survival were also investigated. The suspension of difference in demographic (race), various treatment options, types of chemotherapy and anemic status were analyzed accordingly. Next, the same variables were analyzed in subgroups of P16− and P16+ to investigate any differences in the subgroups. Survival graphs were generated for each step mentioned above.

Following, the significant variables from the complete model selections were analyzed via various distributions in AFT model, which included Gamma, Weibull, exponential, L-normal, and L-logistic. Results and fit statistics were analyzed, and the best models were chosen for the complete survival model. This was achieved by using Proc Lifereg.

All analysis was considered at the significance level of $P \leq 0.05$, unless noted otherwise.

Chapter 3: Results

3.1 Cox Proportional Hazard Model

During the initial data exploration, the full model with all the terms was developed using Cox proportional hazard model. During the process, there were 55 total events which corresponded to the number of patients that were deceased and not censored. Event used and censoring statistics can be seen below:

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
173	55	118	68.21

Table 7

3.1.1 Complete Model

Parameter estimates and Hazard ratios were generated and shown beneath:

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
P16	1	-1.288	0.422	9.311	0.002	0.276
Radmodality	1	-0.025	0.107	0.053	0.818	0.976
LN_positive	1	0.059	0.029	4.235	0.040	1.061
Treatment5	1	0.069	0.177	0.149	0.699	1.071
Age	1	0.023	0.020	1.287	0.257	1.023
Sex	1	-0.617	0.598	1.062	0.303	0.540
Anemic	1	0.011	0.586	0.000	0.985	1.011
Hematocrit	1	0.204	0.220	0.857	0.355	1.226
HGB	1	-0.599	0.644	0.865	0.352	0.549
Chem3	1	-0.208	0.190	1.204	0.273	0.812
Tobacco3	1	0.288	0.206	1.954	0.162	1.333
Alcohol	1	-0.440	0.384	1.311	0.252	0.644
Recurrence	1	1.152	0.818	1.984	0.159	3.166
Locoreg_fail	1	-0.747	0.673	1.233	0.267	0.474

Cancerstatus	1	1.085	0.855	1.610	0.205	2.960
ACE_27	1	0.036	0.177	0.042	0.838	1.037
Differentiation	1	0.481	0.256	3.513	0.061	1.617
Race	1	-0.720	0.585	1.511	0.219	0.487

Table 8

Likelihood ratio, score and Wald tests were performed for the whole model and all are shown significance:

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	70.662	18	<.0001
Score	108.526	18	<.0001
Wald	70.429	18	<.0001

Table 9

3.1.2 Backward Elimination:

Next, stepwise backward selections of the whole model, P16+ status, and P16− status were generated at significance level of $p \leq 0.15$. The predictor with the highest P-value was removed one at the time until all of the parameters had at least $p \leq 0.15$.

Whole model:

The steps of removal are shown in a chart below:

Summary of Backward Elimination					
Step	**Effect Removed**	**DF**	**Number In**	**Wald Chi-Square**	**Pr > ChiSq**
1	Anemic	1	17	0.000	0.985
2	ACE_27	1	16	0.042	0.838
3	Radmodality	1	15	0.054	0.816
4	Treatment5	1	14	0.150	0.699
5	Hematocrit	1	13	0.869	0.351
6	HGB	1	12	0.047	0.829
7	Race	1	11	0.786	0.375
8	Chem3	1	10	0.642	0.423
9	Alcohol	1	9	0.863	0.353
10	Sex	1	8	1.354	0.245

11	Locoreg_fail	1	7	1.285	0.257
12	Recurrence	1	6	1.386	0.239

Table 10

After each parameter was removed, the chi-squares score and corresponding P-values were adjusted to fit the new model. Beneath is an output chart of the remaining parameters after the removal of insignificant factors:

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
P16	1	-1.019	0.332	9.435	0.002	0.361
LN_positive	1	0.043	0.027	2.591	0.108	1.044
Age	1	0.026	0.016	2.496	0.114	1.026
Tobacco3	1	0.278	0.185	2.270	0.132	1.321
Cancerstatus	1	1.906	0.326	34.244	<.0001	6.725
Differentiation	1	0.420	0.235	3.190	0.074	1.522

Table 11

Also, likelihood ratio, score and Wald test scores are significant for the optimized model and shown as:

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	63.030	6	<.0001
Score	101.906	6	<.0001
Wald	69.812	6	<.0001

Table 12

P16+ Status:

For P16+ status with 6 degrees of freedom, likelihood ratio, score, and Wald tests were all significant at <.0001 and had Chi-squared score of 49.019, 88.432, and 53.358, respectively. Eleven predictors which failed to meet the criteria, were removed and the steps of elimination are in the order of Hematocrit, Anemic, Sex, Radmodality, ACE_27, Chem3, Treatment5, Age, Recurrence, Locoreg_fail, and Race. Six predictors remained in the model and seen in Table 13:

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
LN_positive	1	0.059	0.027	4.644	0.031	1.061
HGB	1	-0.244	0.154	2.499	0.114	0.784
Tobacco3	1	0.482	0.206	5.455	0.020	1.619
Alcohol	1	-0.813	0.398	4.180	0.041	0.443
Cancerstatus	1	2.872	0.438	42.918	<.0001	17.673
Differentiation	1	0.533	0.262	4.129	0.042	1.704

Table 13

P16− Status:

With 7 degrees of freedom, the likelihood, score and Wald test for P16− status had Chi-square of 18.534, 12.725, and 11.144 with P-value of 0.005, 0.0476 and 0.084, respectively. Eleven predictors were removed and in the order of Cancerstatus, Race, Sex, LN_positive, Age, Differentiation, Treatment5, Recurrence, Alcohol, ACE_27, and Anemic. The six remaining significant factors are seen beneath:

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
Radmodality	1	0.859	0.322	7.101	0.008	2.360
Hematocrit	1	-0.945	0.435	4.723	0.030	0.389
HGB	1	2.786	1.287	4.682	0.031	16.214
Chem3	1	-1.628	0.526	9.579	0.002	0.196
Tobacco3	1	-1.556	0.602	6.673	0.010	0.211
Locoreg_fail	1	-2.043	1.016	4.041	0.044	0.130

Table 14

3.1.3 Forward Selection

Stepwise forward selection was performed with restriction of $p \leq 0.20$. Starting with the most significant factor, the model was built based on including the parameter with the lowest P-value one at a time till all the p-values with $p \leq 0.20$ were included. Similar to backward elimination, when one new factor is included into the model, each parameter adjusts its Chi-squared score and

corresponding P-value accordingly to fit the new model. Forward selection models were generated for the complete model, P16+ and P16−.

Whole model:

In the complete model selection, likelihood, score and Wald tests all were significant at 0.05 level and the results are show below:

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	67.350	6	<.0001
Score	103.019	6	<.0001
Wald	73.019	6	<.0001

Table 15

Six predictors were selected at p≤0.20, and the other six were disregarded. The output of the parameters and the order of entry for each factor are presented below:

Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**	**Order of Entry**
P16	1	-1.019	0.332	9.435	0.002	0.361	2
LN_positive	1	0.043	0.027	2.591	0.108	1.044	6
Age	1	0.026	0.016	2.496	0.114	1.026	5
Tobacco3	1	0.278	0.185	2.270	0.132	1.321	3
Cancerstatus	1	1.906	0.326	34.244	<.0001	6.725	1
Differentiation	1	0.420	0.235	3.190	0.074	1.522	4

Table 16

1) P16+ Status

For P16+ population with 6 degree of freedom and P-value of <0.001, the likelihood, score and Wald test had results of 49.019, 88.432, and 53.358, respectively. Result for this model is shown below:

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**	**Order of Entry**

LN_positive	1	0.059	0.027	4.644	0.0312	1.061	5
HGB	1	-0.244	0.154	2.499	0.114	0.784	6
Tobacco3	1	0.482	0.206	5.455	0.0195	1.619	2
Alcohol	1	-0.813	0.398	4.180	0.0409	0.443	4
Cancerstatus	1	2.872	0.438	42.918	<.0001	17.673	1
Differentiation	1	0.533	0.262	4.129	0.0421	1.704	3

Table 17

P16− status

For P16− population, with 6 degree of freedom and P-value of 0.0063, 0.0136, and 0.0444, the likelihood ratio, score and Wald test had results of 17.984, 16.027, and 12.918 respectively. The output for parameter estimates and order of entry is show beneath:

Analysis of Maximum Likelihood Estimates							
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**	**Order of Entry**
Radmodality	1	0.356	0.190	3.505	0.061	1.427	3
Treatment5	1	0.717	0.405	3.129	0.077	2.048	5
Anemic	1	-1.061	0.738	2.065	0.151	0.346	6
Chem3	1	-1.293	0.472	7.510	0.006	0.274	1
Cancerstatus	1	1.750	0.760	5.301	0.021	5.757	2
Race	1	-1.795	0.743	5.835	0.016	0.166	4

Table 18

3.1.4 Class P16

Additionally, P16 was treated as a class within Proc Phreg and a backward elimination at P≤0.20 level. With 4 degree of freedom the likelihood ratio, score and Wald test results are 53.868, 75.277, and 60.431 with P-value <0.0001. The Parameter estimate is presented below:

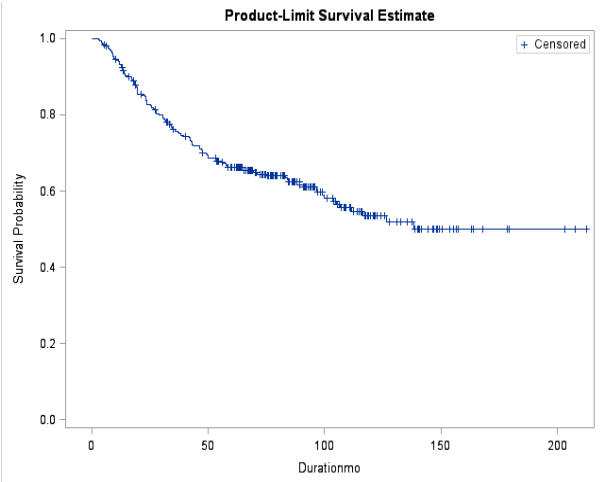
Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
Age	1	0.026	0.016	2.554	0.110	1.026
Tobacco3	1	0.315	0.179	3.107	0.078	1.371
Recurrence	1	1.878	0.294	40.784	<.0001	6.540
Differentiation	1	0.692	0.207	11.168	0.001	1.999

Table 19

3.2 Kaplan-Meier Method

3.2.1 Whole Model

First, survival graphs were generated for HNSCC as a whole with Durationmo as the response variable and Vitalstatus=0 as the censoring. The graphs are presented below:



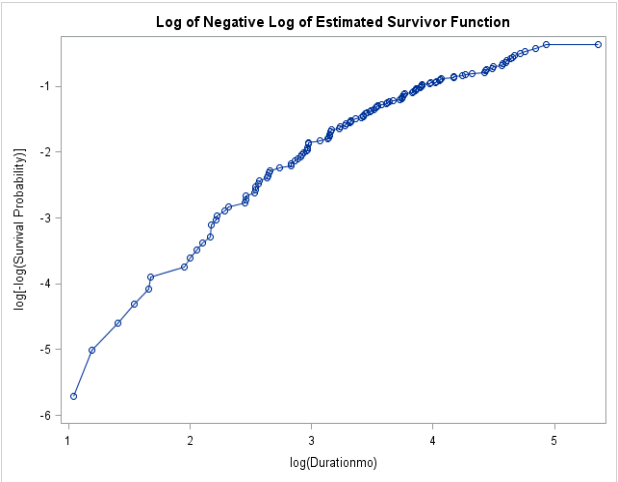


Figure 4

To validate the backward and forward selection in Proc Phreg, test statement was used for Proc Lifetest which produced a composite of univariate estimate with standard deviation and corresponding chi-squared statistics, as well as forward selection of the whole model. This syntax is optimal with binary data such as many variables presented in this dataset. The univariate statistics with all parameters is described in table 20:

Univariate Chi-Squares								
	Log-Rank Test				Wilcoxon Test			
Variable	**Test Stats**	**SE**	**Chi-Square**	**Pr >Chi**	**Test Stats**	**SE**	**Chi-Square**	**Pr >Chi**
P16	12.55	2.20	32.57	<.0001	10.59	2.18	23.63	<.0001
Radmodality	-7.79	10.90	0.51	0.475	-5.33	9.18	0.34	0.561
LN_positive	-10.07	33.33	0.09	0.763	-11.66	27.91	0.17	0.676
Treatment5	14.31	6.93	4.26	0.039	12.46	6.02	4.28	0.039
Age	-128.6	63.71	4.07	0.044	-98.35	53.89	3.33	0.068

Sex	1.69	2.42	0.49	0.486	1.76	2.02	0.76	0.384
Anemic	-2.27	2.58	0.78	0.377	-1.95	2.20	0.79	0.375
Hematocrit	6.76	25.02	0.07	0.787	3.02	21.31	0.02	0.887
HGB	5.47	9.00	0.37	0.544	3.51	7.63	0.21	0.646
Chem3	13.36	7.27	3.38	0.066	11.35	6.10	3.46	0.063
Tobacco3	-16.18	6.40	6.39	0.012	-13.05	5.24	6.21	0.013
Alcohol	-1.93	3.36	0.33	0.565	-2.16	2.80	0.60	0.439
Recurrence	-15.21	2.08	53.40	<.0001	-12.83	2.14	35.84	<.0001
Locoreg_fail	-1.55	1.19	1.69	0.193	-1.25	1.12	1.26	0.262
Cancerstatus	-14.83	1.71	75.08	<.0001	-12.75	1.90	45.03	<.0001
ACE_27	-9.89	6.14	2.60	0.107	-8.49	5.26	2.61	0.106
Differentiation	-17.92	4.83	13.75	0.000	-14.89	4.11	13.14	0.000
Race	-2.69	1.95	1.90	0.168	-2.14	1.73	1.53	0.216

Table 20

The forward selection produced by Proc Lifetest can be shown as:

Forward Stepwise Sequence of Chi-Squares							
		Log-Rank Test			**Wilcoxon Test**		
DF	**Pr >Chi**	**Variable**	**Chi Increment**	**Pr> Increment**	**Variable**	**Chi Increment**	**Pr> Increment**
1	<.0001	**Cancerstatus**	75.08	<.0001	**Cancerstatus**	45.03	<.0001
2	<.0001	**P16**	16.98	<.0001	**P16**	12.34	0.0004
3	<.0001	**Tobacco3**	4.172	0.041	**Tobacco3**	5.176	0.023
4	<.0001	**Differentiation**	3.107	0.078	**Locoreg_fail**	2.464	0.117
5	<.0001	**Locoreg_fail**	2.205	0.138	**Differentiation**	1.797	0.180
6	<.0001	**Recurrence**	2.791	0.095	**Recurrence**	1.848	0.174
7	<.0001	**Age**	1.253	0.263	**LN_positive**	1.300	0.254
8	<.0001	**LN_positive**	1.126	0.289	**Age**	1.226	0.268
9	<.0001	**Alcohol**	0.703	0.402	**Chem3**	0.457	0.499
10	<.0001	**ACE_27**	0.266	0.606	**Alcohol**	0.361	0.548
11	<.0001	**Chem3**	0.248	0.618	**Sex**	0.305	0.581
12	<.0001	**Race**	0.151	0.698	**Race**	0.214	0.644
13	<.0001	**Anemic**	0.071	0.790	**ACE_27**	0.275	0.600
14	<.0001	**Hematocrit**	0.034	0.853	**Radmodality**	0.193	0.660
15	<.0001	**HGB**	0.309	0.578	**Anemic**	0.150	0.699
16	<.0001	**Sex**	0.015	0.902	**Treatment5**	0.123	0.725
17	<.0001	**Radmodality**	0.006	0.938	**HGB**	0.010	0.922
18	<.0001	**Treatment5**	0.002	0.964	**Hematocrit**	0.060	0.807

Table 21

The significant variables that are ≤0.20 for both tests are Cancerstatus, P16, Tobacco3, Differentiation and Locoreg_fail.

3.2.2 Strata P16

Next, P16 was treated as strata by itself. The two groups had significantly different survival rate, which is presented in table 21:

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	68.48	1	<.0001
Wilcoxon	64.86	1	<.0001
-2Log(LR)	55.49	1	<.0001

Table 22

Survival graph for P16+ and P16− were generated and shown in figure 5:

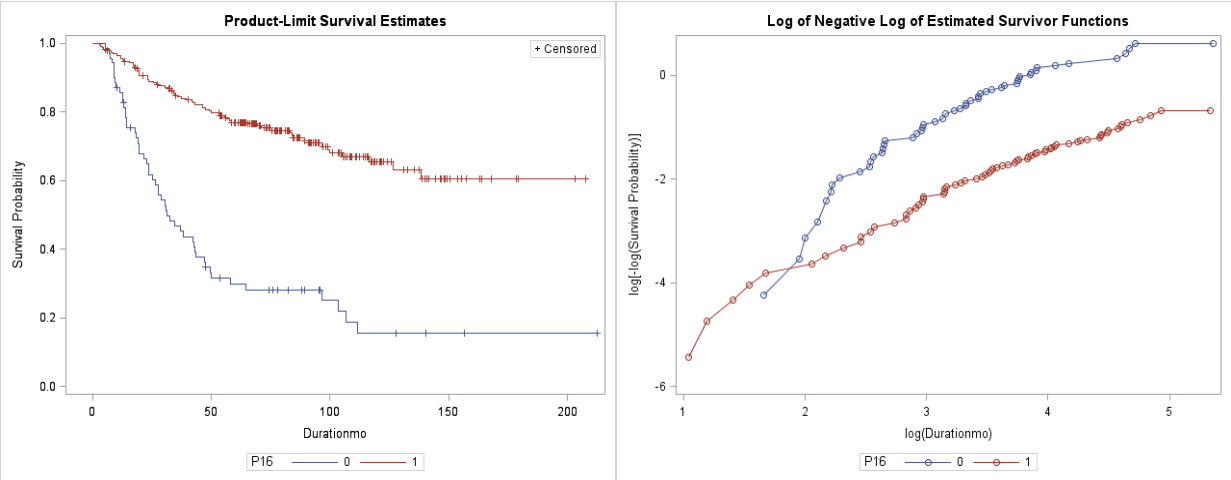


Figure 5

P16+ status has better survival rate than P16− patients for this dataset.

To confirm results from the class statement of Cox proportional hazard model, test was used again with P16 desinated as strata, and result as followed:

Univariate Chi-Squares		
	Log-Rank Test	**Wilcoxon Test**

Variable	Test Stats	SD	Chi-Square	Pr> Chi	Test Stats	SD	Chi-Square	Pr >Chi
Radmodality	-4.49	10.77	0.174	0.677	-0.26	8.742	0.001	0.976
LN_positive	-25.2	30.39	0.685	0.408	-27.76	25.55	1.180	0.277
Treatment5	8.096	7.346	1.214	0.271	5.667	5.678	0.996	0.318
Age	-115.0	69.65	2.725	0.099	-96.62	51.97	3.457	0.063
Sex	1.356	2.364	0.329	0.566	1.302	1.937	0.452	0.501
Anemic	-0.501	2.746	0.033	0.855	-0.90	2.119	0.180	0.672
Hematocrit	-5.64	26.36	0.046	0.831	-2.828	20.60	0.019	0.891
HGB	0.002	9.333	0.000	1.000	0.474	7.343	0.004	0.949
Chem3	8.125	7.080	1.317	0.251	4.423	5.621	0.619	0.431
Tobacco3	-12.85	6.145	4.374	0.037	-11.27	4.958	5.169	0.023
Alcohol	0.922	3.050	0.091	0.762	0.167	2.596	0.004	0.949
Recurrence	-12.78	2.373	29.023	<.0001	-10.54	2.088	25.51	<.0001
Locoreg_fail	-0.105	1.591	0.004	0.948	-0.07	1.156	0.003	0.955
Cancerstatus	-12.97	2.028	40.925	<.0001	-10.57	1.885	31.423	<.0001
ACE_27	-6.39	6.712	0.907	0.341	-5.64	5.035	1.257	0.262
Differentiation	-10.25	4.399	5.432	0.020	-8.103	3.648	4.933	0.026
Race	0.687	2.294	0.090	0.765	0.521	1.665	0.098	0.754

Table 23

Table of forward selection from Proc Lifetest is as followed:

Forward Stepwise Sequence of Chi-Squares							
		Log-Rank Test			**Wilcoxon test**		
DF	**Pr> Chi**	**Variable**	**Chi Increment**	**Pr> Increment**	**Variable**	**Chi Increment**	**Pr> Increment**
1	<.0001	**Cancerstatus**	40.92	<.0001	**Cancerstatus**	31.42	<.0001
2	<.0001	**Tobacco3**	6.714	0.010	**Tobacco3**	6.107	0.014
3	<.0001	**Differentiation**	3.072	0.080	**Differentiation**	2.697	0.101
4	<.0001	**Age**	2.007	0.157	**Locoreg_fail**	1.678	0.195
5	<.0001	**Recurrence**	2.243	0.134	**Recurrence**	1.990	0.158
6	<.0001	**Locoreg_fail**	3.103	0.078	**Age**	1.591	0.207
7	<.0001	**LN_positive**	2.510	0.113	**LN_positive**	1.916	0.166
8	<.0001	**Treatment5**	1.168	0.280	**Race**	0.802	0.371
9	<.0001	**Chem3**	1.775	0.183	**Sex**	0.659	0.417
10	<.0001	**Race**	0.506	0.477	**Chem3**	0.480	0.488
11	<.0001	**Sex**	0.341	0.559	**Treatment5**	0.426	0.514
12	<.0001	**Alcohol**	0.294	0.588	**Alcohol**	0.343	0.558

13	<.0001	**Radmodality**	0.299	0.585	**Radmodality**	0.435	0.510
14	<.0001	**Anemic**	0.062	0.803	**ACE_27**	0.385	0.535
15	<.0001	**Hematocrit**	0.198	0.657	**Anemic**	0.335	0.563
16	<.0001	**HGB**	0.425	0.514	**Hematocrit**	0.007	0.935
17	<.0001	**ACE_27**	0.0002	0.988	**HGB**	0.214	0.644

Table 24

The significant variables ($p \leq 0.20$) from forward selection were Cancerstatus, Tobacco3, Differentiation, Recurrence, Age, Locoreg_fail and LN_positive. Chem3 was positive for the log rank test and Age had slightly larger P-value than 0.20 in the Wilcoxon test.

Following, P16 was treated as strata with other factors including Age, Tobacco3, Recurrence, Differentiation, Locoreg_fail and LN_positive. The chosen factors were the significance variables in the previous model. Age was grouped into 2 categories, <55 and ≥55. LN_positive was grouped into >10, between 10 and 20, and ≥20. The Log-Rank, Wilcoxon, and −2Log (LR) test results all had $p<0.0001$ for all groups and table for Chi-squared is presented below.

Test of Equality over Strata (Chi-Square Test)						
Variables	Age	Tobacco3	Recurrence	Locoreg _fail	Differentiation	LN_ Positive
Log-Rank	74.30	72.59	91.56	66.75	90.45	66.26
Wilcoxon	69.89	65.88	84.90	63.85	79.81	63.32
-2Log(LR)	64.52	69.87	68.49	54.17	71.84	47.38

Table 25

The Survival probability graphs can be seen below:

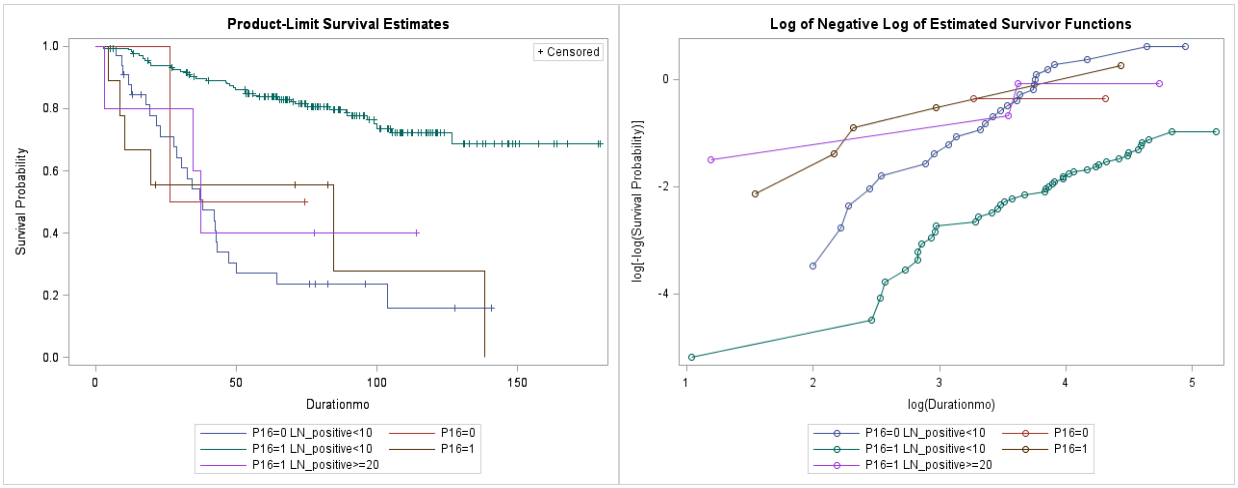


Figure 6

For lymph nodes that were positive for cancer, the best survival rate group was the P16+ and lymph node found less than 10. Others are similar and some had not enough data points to determine.

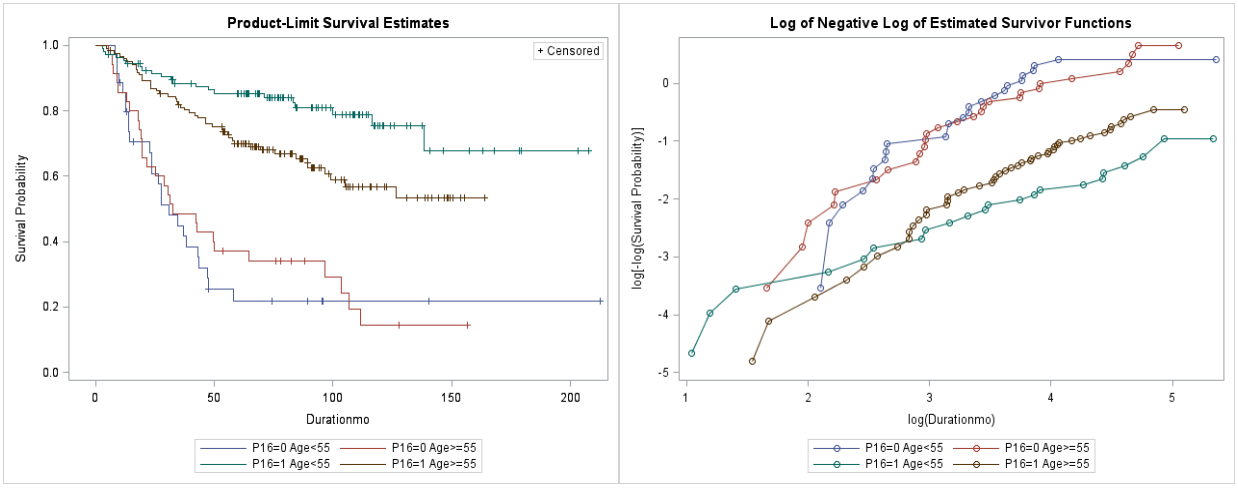


Figure 7

In these strata, P16+ and less than 55 years old group had the best survival rate followed by P16+ and older than 55 years old group. The rest are simiar in survival rate.

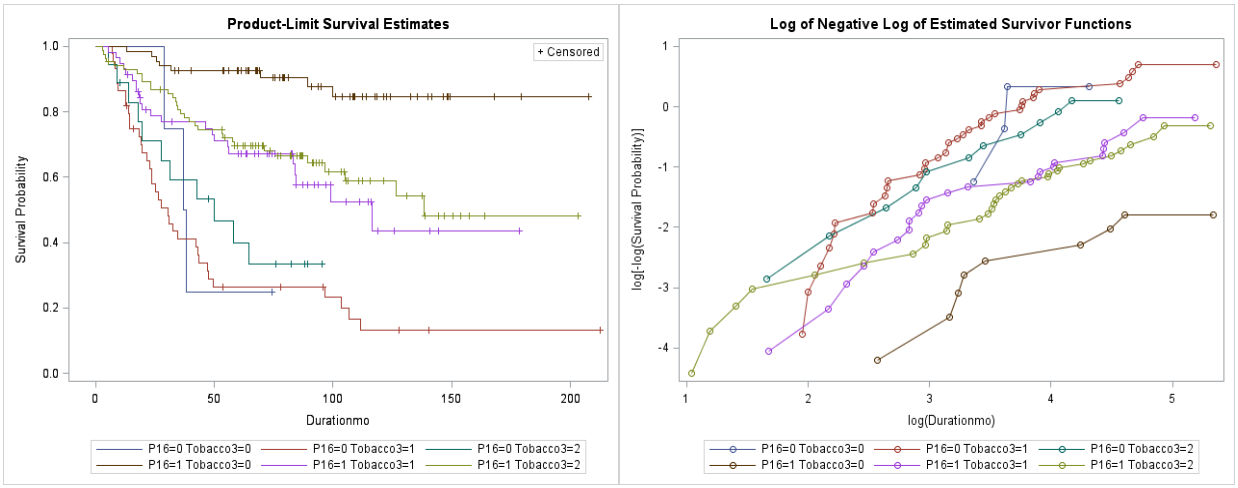


Figure 8

For tobacco status, P16+ / nonsmokers have the best survival rate, Followed by P16+ / former smoker and P16+ / current smoker. P16− / current smoker had the worst survival rate.

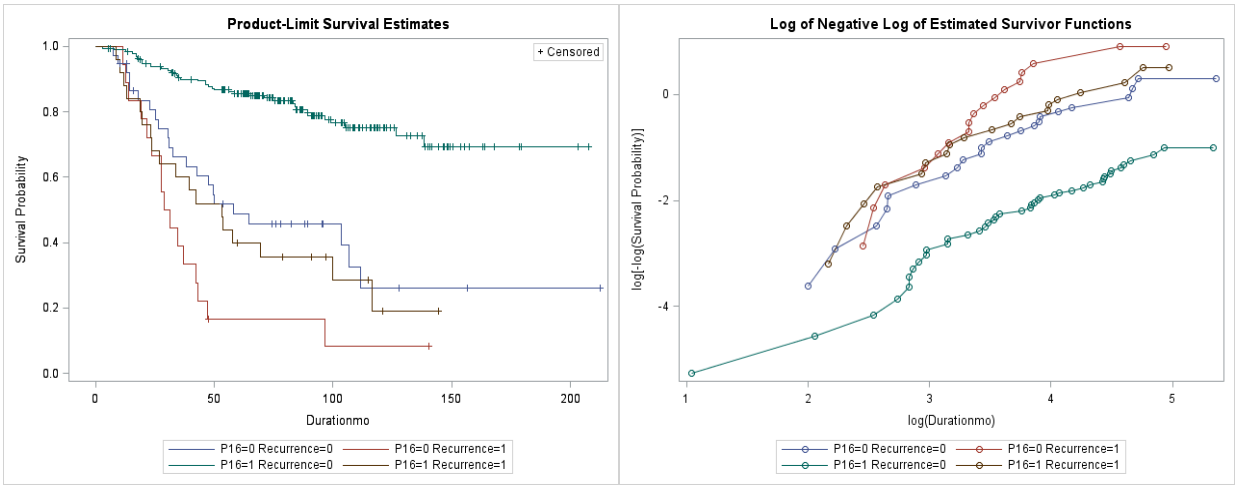


Figure 9

In this strata, P16+ / no recurrence of cancer had significantly better survival rate, followed by P16− / no recurrence group, then by P16+ / recurrence, and lastly is the P16− / recurrence group.

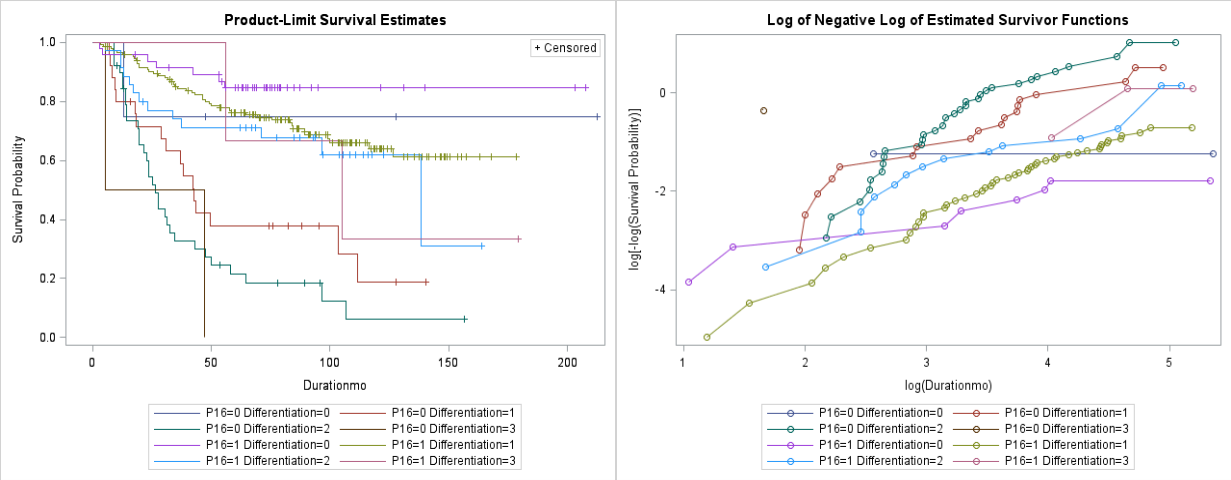


Figure 10

In this model, many did not have enough data to be accurate, such as the P16− / no differentiation, P16− / well differentiated and P16+ / well differentiated groups. The best survival rate group is P16+ / no differentiation followed by P16+ / poorly differentiated, then by P16+ / moderately differentiated, P16− / poorly differentiated and last, P16− / moderately differentiated.

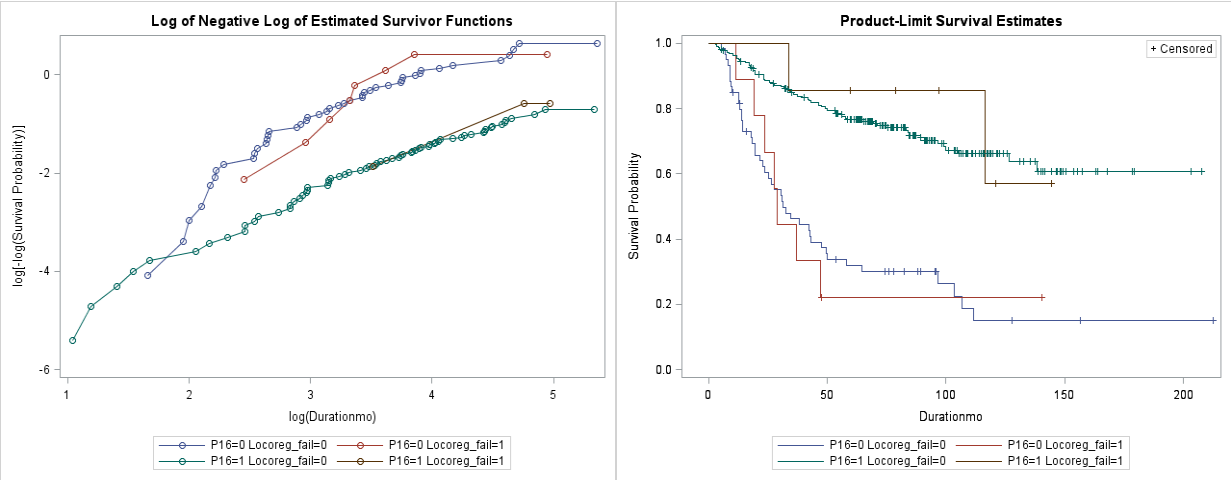


Figure 11

In this group, P16+/ has loco-regional failure and P16−/has loco-regional failure did not have as many data point. The survival rate seems to be separated by P16 status.

3.2.3 Other Factors

Few variables of interest that were not selected from forward selection of Proc Lifetest were treated as strata to see if there were differences within strata. The reason for choosing these variables were to explored the effectiveness of different chemotherapy (Chem3), especially since Chem3 had P-value<0.20 in log rank test during forward selection of P16 strata, types of treatment (Treatment5), and difference in race are often questioned. Also, as mentioned in the introduction, anemia is more prevalent in P16− group. The next question is that if anemia plays a role in survival and the problem is investigated subsequently.

First Treatment5 was analyzed and test scores are presented below.

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	43.700	4	<.0001
Wilcoxon	51.017	4	<.0001
-2Log(LR)	35.741	4	<.0001

Table 26

Treatment stratified survival graphs are shown as:

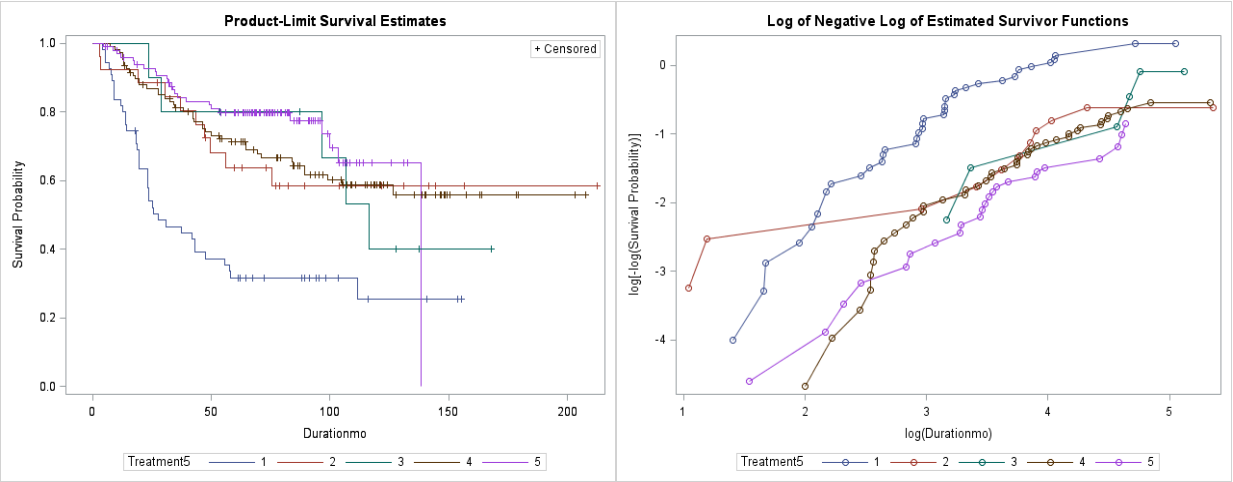


Figure 12

Treatment5 strata were significantly different, with stratum 1 (no treatment) had the lowest survival rate. Stratum 3 (surgery only) and 5 had the best survival rate and were similar to each other. Rest of the two strata were in the middle and were similar.

Next, Chem3 was treated as strata and the result is present as:

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	24.214	2	<.0001
Wilcoxon	30.397	2	<.0001
-2Log(LR)	18.821	2	<.0001

Table 27

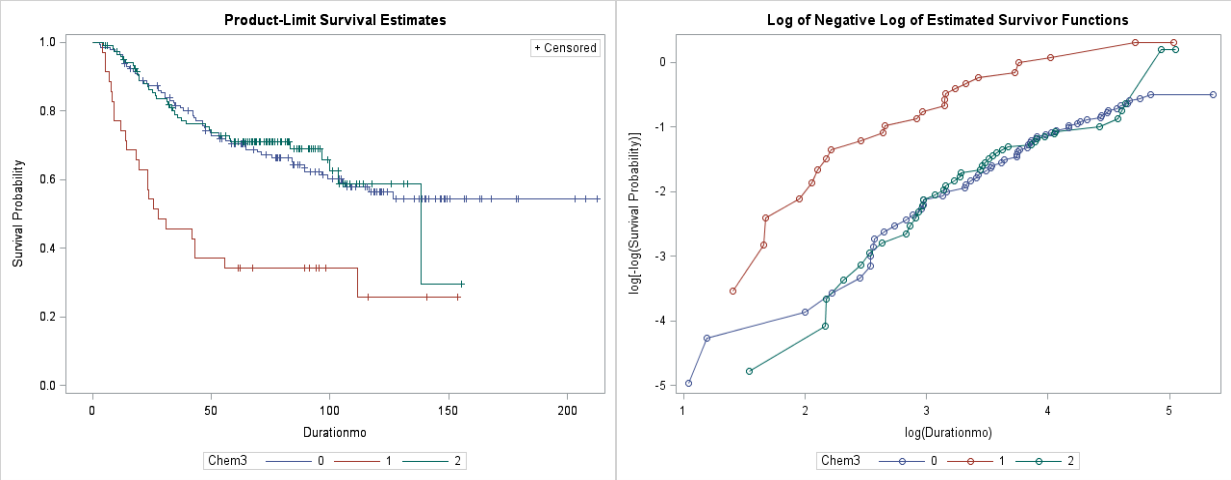


Figure 13

Chem3 showed significant difference and all strata are not equal. Stratum Chem3=1 had the lowest survival rate which is induction chemotherapy. The other 2 strata showed similarity in survival probability, which were no chemotherapy and concurrent chemotherapy.

Furthermore, Race was analyzed and hypothesis test statistics are below:

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	17.791	2	0.0001
Wilcoxon	15.976	2	0.0003
-2Log(LR)*	16.541	2	0.0003

Table 28

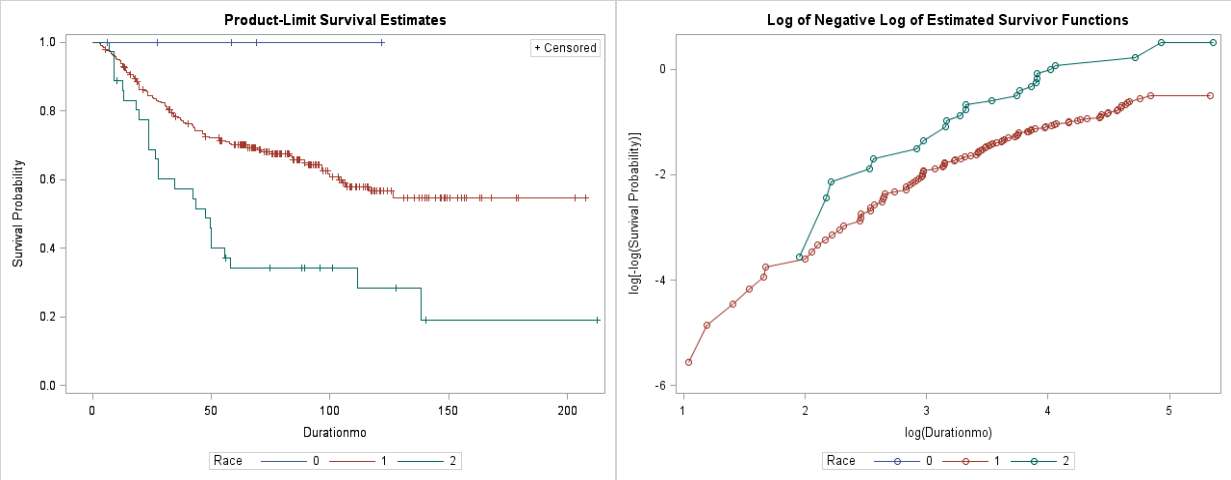


Figure 14

In stratum 0 which is the category of other, did not have enough data point and resulted in an empty stradum. As a reslut, -2Log(LR) statistics might not be correct since it assume proportionality. Ohterwise, the strata show significant in difference on the 0.05 level. Additionally, P16+ and P16− status were treated separately for the variables analyzed in the previous section, which include Treatment5, Chem3 and Race. A table of P16+ group is presented beneath:

Test of Equality over Strata (P16+)									
	Treatment5			Chem3			Race		
Test	**Chi-Square**	**DF**	**Pr >Chi-Square**	**Chi-Square**	**DF**	**Pr >Chi-Square**	**Chi-Square**	**DF**	**Pr >Chi-Square**
Log-Rank	15.395	4	0.004	10.8	2	0.005	3.135	2	0.209
Wilcoxon	21.121	4	0.000	15.5	2	0.000	1.720	2	0.423
-2Log(LR)	12.357	4	0.015	8.5	2	0.014	3.989	2	0.136

Table 29

Survival graphs for all factors are seen below:

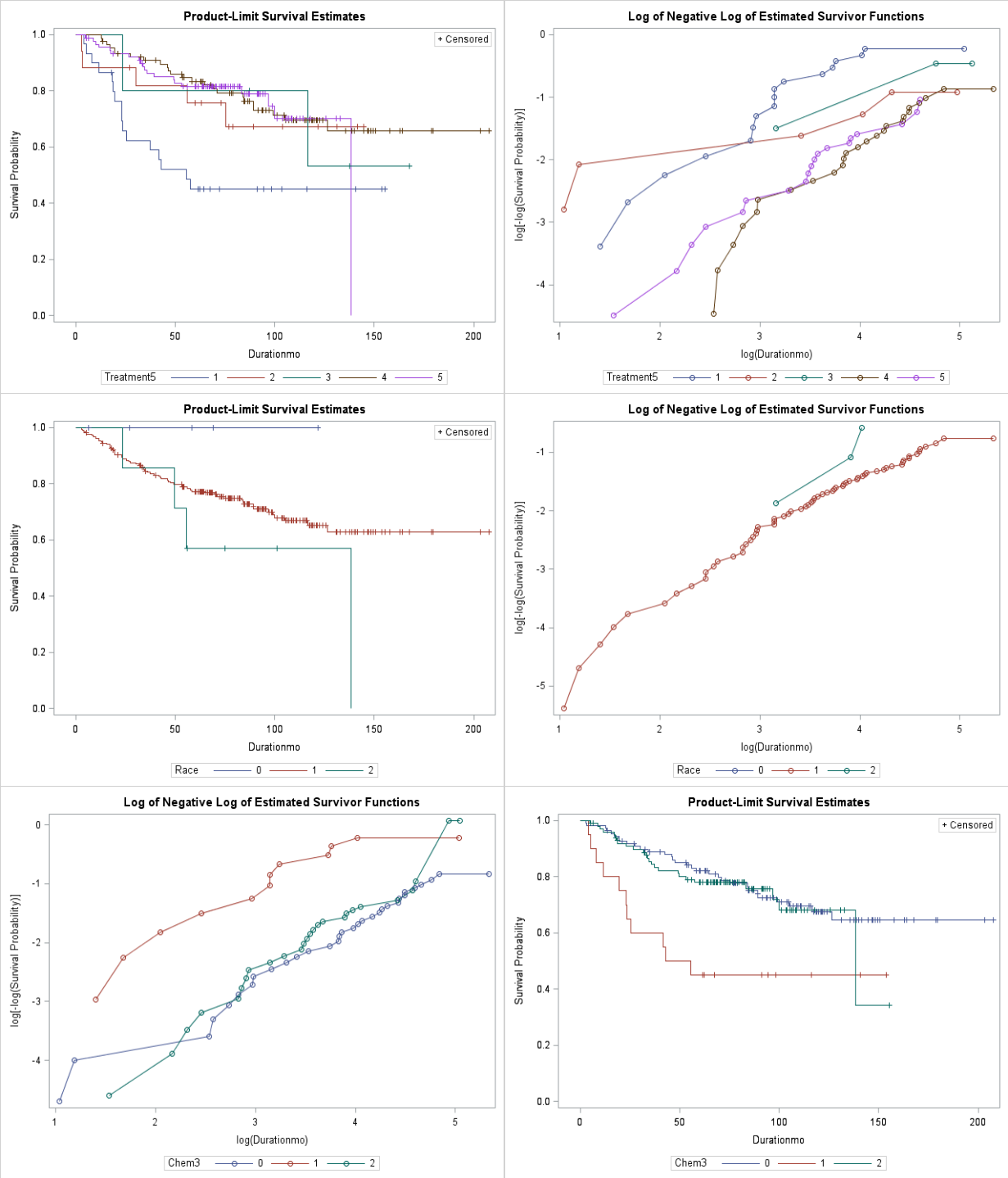


Figure 15

For P16+ subgroup, Race was the only factor that did not show significance which could be the result of not having enough data points for the analysis. No treatment group had the worst survival outcome with rest of the strata appeared to have similar survival rate. Induction

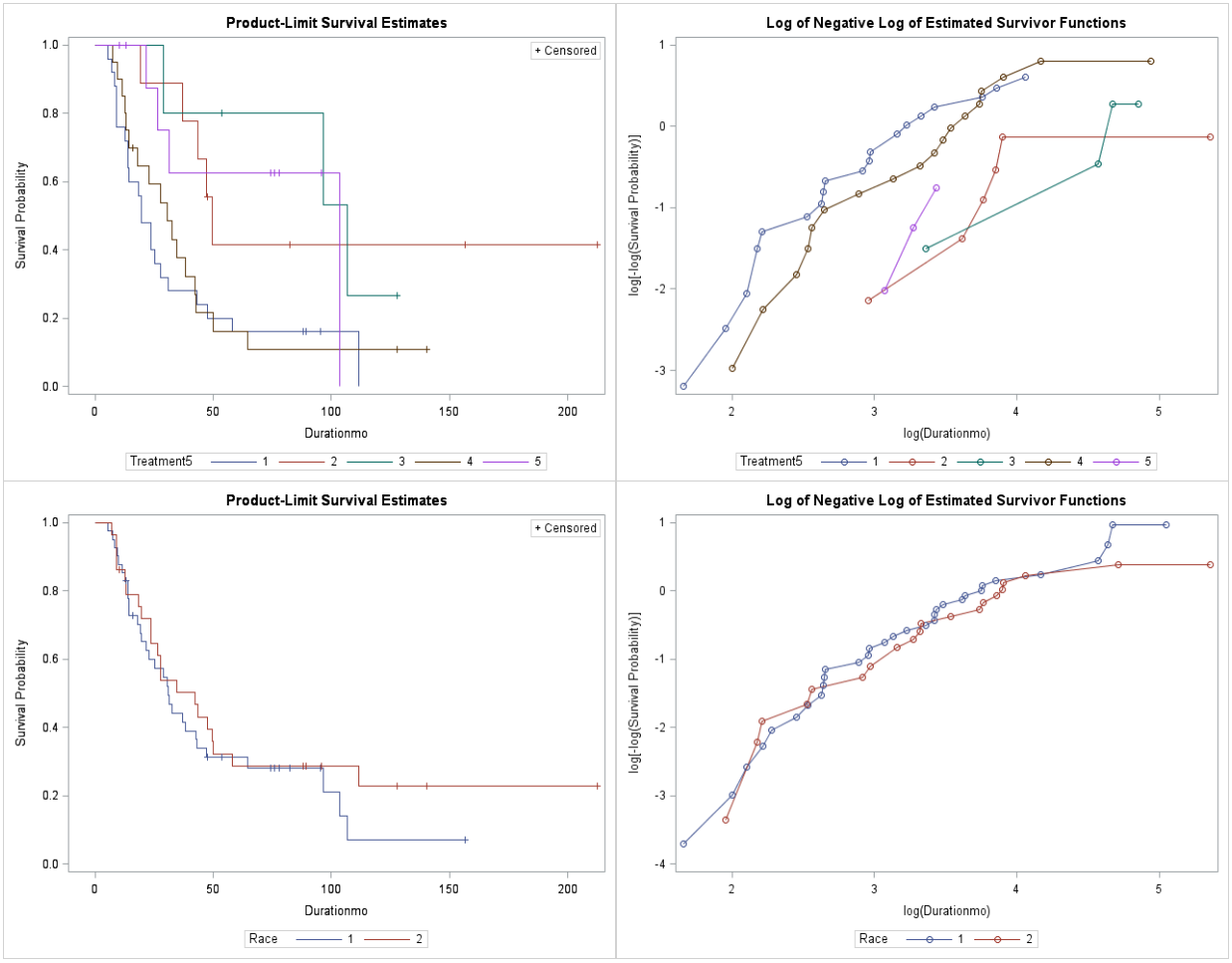
chemotherapy had the worst survival rates which were similar to when chemotherapy were treated as strata without consideration of P16.

Following are the test statistics for P16− group:

Test of Equality over Strata (P16−)									
	Treatment5			Chem3			Race		
Test	**Chi-Square**	**DF**	**Pr >Chi-Square**	**Chi-Square**	**DF**	**Pr >Chi-Square**	**Chi-Square**	**DF**	**Pr >Chi-Square**
Log-Rank	13.599	4	0.009	3.963	2	0.138	0.644	1	0.422
Wilcoxon	16.421	4	0.003	6.337	2	0.042	0.251	1	0.617
-2Log(LR)	16.401	4	0.003	3.352	2	0.187	1.271	1	0.260

Table 30

Survival graphs of each variable are as presented:



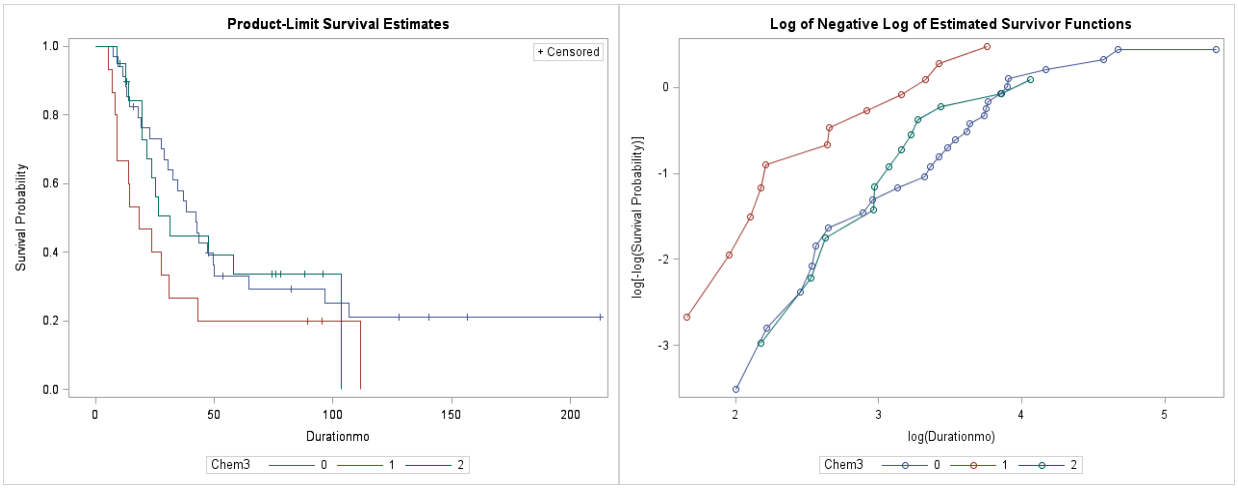


Figure 16

For P16− group, treatment5 showed most significance in difference of strata, followed by Chem3. Race was not significantly different for this group. Chemotherapy result is similar to P16+ group. For treatment, no treatment received had the worst survival rate for both with P16+ group had more difference.

Lastly, Anemic was analyzed. First, anemic was treated as strata alone and the result is as followed:

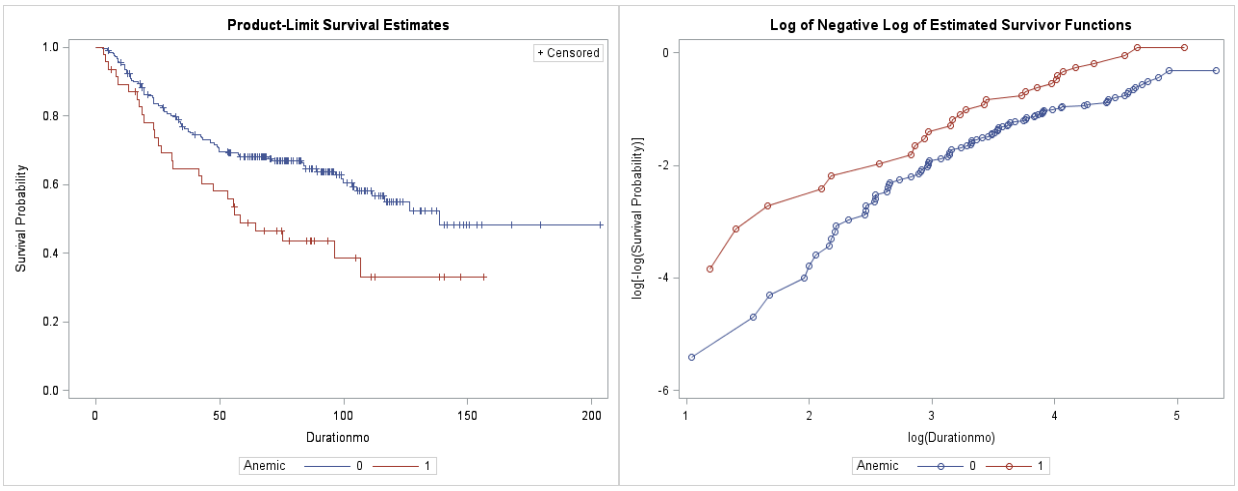


Figure 17

the log-rank, wilcoxon and -2log(LR) were all significant on the 0.05 level. Patients that were not anemic havd better survival rate.

Next, Anemia was analyzed against P16 status via strata as well and the log-rank, wilcoxon and -2log(LR) were all significant with P<0.0001. the results are as presented:

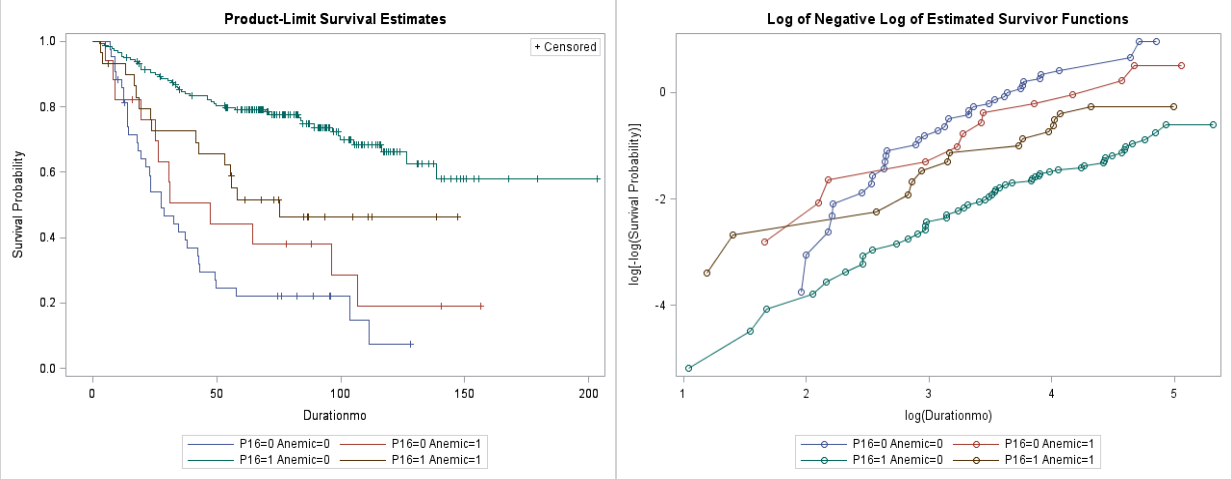


Figure 18

For P16+ group, result as expected where none anemic group had better survival rate. However, for this dataset, patients who are P16−/ anemic, seemed to have better survival rate after 20 months. This was further investigated where P16− status was specified and here is the result:

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	1.35	1	0.246
Wilcoxon	0.93	1	0.335
-2Log(LR)	2.14	1	0.144

Table 31

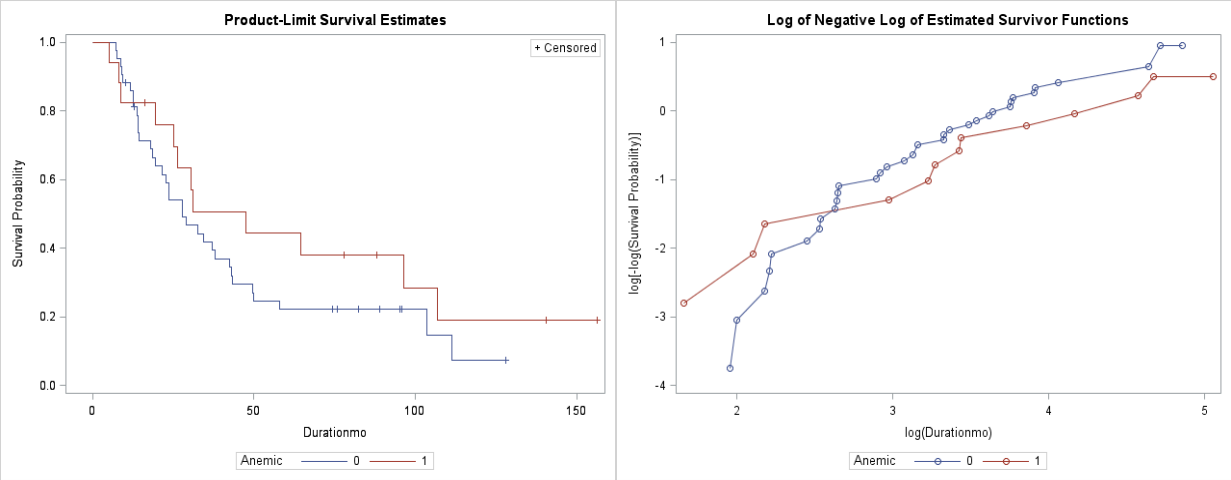


Figure 19

There was a small difference but not significant at the 0.05 level. Anemic+ group seems to have slight better survival rate.

3.3 Accelerated Fail Time Model

3.3.1 Whole model with various distributions

Various Factors that showed significance in previous models selections were considered, and after some analyze, six were used to fit the different distribution in AFT models. The variable chosen are P16, LN_positive, Age, Recurrence, Locoreg_fail, and Differentiation. The censoring statistics is shown as:

Number of Observations Read	300
Number of Observations Used	222
Missing Values	78

Table 32

All parameters were significant at P≤0.05 level in chi-squared test, and the parameter estimate can be seen in table 21:

Analysis of Maximum Likelihood Parameter Estimates						
		Gamma	**Weibull**	**Exponential**	**L-normal**	**L-logistic**
Parameter	**DF**	**Estimate**	**Estimate**	**Estimate**	**Estimate**	**Estimate**
Intercept	1	6.910	7.640	8.490	6.762	6.933

P16	1	0.902	0.969	1.138	1.194	1.086
LN_positive	1	-0.039	-0.062	-0.075	-0.062	-0.063
Age	1	-0.025	-0.039	-0.049	-0.034	-0.035
Recurrence	1	-1.219	-1.584	-1.865	-1.616	-1.577
Locoreg_fail	1	0.787	0.792	0.951	1.103	0.945
Differentiation	1	-0.386	-0.393	-0.484	-0.301	-0.328
Scale	1	0.103	0.763	1.000	1.226	0.647
Shape	1	10.628	1.311	1.000		

Table 33

3.3.2 Fit Statistics comparison

Fit Statistics (logged response)					
	Gamma	**Weibull**	**Exponential**	**L-normal**	**L-logistic**
-2 Log Likelihood	247.63	322.18	328.34	332.72	327.82
AIC (smaller is better)	265.63	338.18	342.34	348.72	343.82
AICC (smaller is better)	266.48	338.86	342.86	349.40	344.50
BIC (smaller is better)	296.25	365.41	366.16	375.95	371.04

Table 34

Comparing -2 Log likelihood, AIC, AICC and BIC results of the various distributions, Gamma distribution appears to be the most optimal, following by Weibull distribution. Models from those two distributions are the best fit and presented below.

Gamma Distribution Model

$$\log h(t)=6.91_{Intercept}+0.902XP_{P16}-0.039X_{LN\_positive}-0.025XAge-1.219X_{Recurrence}+0.787X_{Locoreg\_fail}-0.386X_{Differentiation}$$ Equation 27

Weibull Distribution Model

$$\log h(t)=7.640_{Intercept}+0.969X_{P16}-0.062X_{LN\_positive}-0.039X_{Age}-1.584X_{Recurrence}+0.792X_{Locoreg\_fail}-0.393X_{Differentiation}$$ Equation 28

Chapter 4: Discussion

Most of the results found in this study agreed with previous findings. Many treatment options, race, and tobacco use did not have enough data points to give a clear result. Especially when interactions with other factors were analyzed, most variables in this dataset have empty or few data points in one of the category or factor level, thus making interactive terms difficult to determine. Interactive terms were either inaccurate or not able to compute. However, this study provided some valuable information regarding HNSCC, and can potentially contribute more with follow up studies.

In Cox proportional hazard method, during backward and forward selection, the significant factors came to be the same, which are P16, LN_positive, Age, Tobacco3, Cancerstatus, and Differentiation. From literature, P16 was found to have influence in survival rate where P16+ patients had better survival rate, which was mentioned in introduction such that P16+ individuals showed better survival rate amongst HNSCC patients. It's intuitive to assume that number of lymph node found positive for cancer would have a relationship with survival. Age has traditionally found to be an influence in any cancer survival. Since younger patients have much better physical health. Moreover, current cancer status (Cancerstatus) is more likely to influence survival rate. Cancer free patients should have better survival rate. Lastly, stages of cancer, which is represented by Differentiation, can definitely play a role on survival rate, where if the cancer is in the further stage, the survival rate may not be as optimal compared with earlier stage. The variable selections for this model appear to be reasonable.

For P16+ populations, which are predominately HPV+ demonstrated by previous studies, the significant factors from backward elimination are LN_positive, HGB, Tobacco3, Alcohol,

Cancerstatus, and Differentiation. The significant factors from forward selection are the same as backward. This is interesting since Tobacco and Alcohol use is typically related to P16− group. However, excess tobacco and alcohol use is ideal under any circumstances. The other factors are typical indications of cancer that were mentioned in the previous paragraph. Also HGB will be discussed in the next paragraph.

For P16− group, backward elimination selected Radmodality, Hematocrit, HGB, Chem3, Tobacco3, and Locoreg_fail, while forward selection gave Radmodality, Treatment5, Anemic, Chem3, Cancerstatus and race. The difference might be that the P-values for backward and forward selections were slightly different, backward was at ≤0.15 and forward was at ≤0.20. Also, the methods which factors are selected are different. For backward elimination, the whole model is considered first. Within the whole model, there might be co-linearity or interaction of the terms that might affect the P-value of a factor, and P-value is the determine criteria for deletion. In forward selection, the most significant factor was included followed by the second and so on. This method cannot take account that the next factor selected is correlated to the other factors thus may not select the best factors for the whole model. For example, HGB has direct relations to anemic and Hematocrit since HGB, which represents hgb and it is a measurement of blood cell count; while anemic is measuring if someone is below the standard red blood cell count. Hematocrit is a measurement of blood cell count as well. With different methods of selection, one may select one instead of the other due to its mechanism. Here, the conclusion is that red blood cell count is related to P16− HNSCC patients, as mentioned in the introduction that P16− group are often anemic. One more interesting find is that hgb (HGB) was found significant for P16+ group as well suggesting that hgb level also plays a role in survival rate of P16+ group. Furthermore, treatment5, Race, Cancerstatus, Locoreg_fail, and Tobacco3 were the

other differences between forward and backward selection. Treatment5, which is the types of treatment, have similar categories as Chem3, which is the type of chemotherapy. Treatment5 was selected 2nd, it is possible that the model selected chem3 subsequently without consider the two might have relationship. While in backward elimination, treatment5 was the 7th to be deleted, Chem3 might affect the P-value of that. Further analysis using Proc Corr confirmed that Chem3 is correlated with treatment5 and have P-value of <0.0001. Also, Cancerstatus showed correlation with treatment as well, which is intuitive. Radmodality is the different types of radiation and correlated with types of treatment. Additionally, the factors that were significant have effects on cancer survival in general or associated with P16 status. Tobacco use is typically associated with P53 HNSCC, which consists of a large percentage of P16− group. Locoregional failure (Locreg_fail) is cancer reappearance after chemotherapy in the local and regional area. Recurrence of cancer definitely decreases survival rate. Lastly, in category Race, most patients are in category 1 which is Caucasian. In subsequent analysis, race was further investigated. When P16 status was treated as a class against other factors and backward elimination was preformed, the significant factors were Age, Tobacco3, Recurrence and Differentiation. This suggests that these factors are significant when considered interaction with P16. This was further validated in Kaplan Meier method.

Using Kaplan Meier method, backward and forward selection was further validated with forward selection in Proc Lifetest. The forward selection was utilized for the whole model as well as defining strata P16. In the whole model, the significant factors were Cancerstatus, P16, Tobacco3, Differentiation and Locoreg_fail. The difference factors between the Proc Phreg and Proc Lifetest were Age and LN_positive which were only in Proc Phreg, and Locroreg_fail which was only in Proc Lifetest. This is due to the difference in methods. Proc Phreg of Cox

proportional hazard method assumes proportionality while Proc Lifetest doesn't. Age and LN_positive might fit proportional criteria, thus selected by Proc Phreg. Also, Kaplan Meier method is univariate, nonparametric test and good for binary data, while Cox proportional hazard method is semi-parametric. This might be the reason Locroreg_fail was selected for Proc Lifetest. Since Locroreg_fail is stored as binary data.

When P16 was treated as strata over event time, P16+ group had much better survival rate, which is confluent with current studies. In the forward selection here, the significant factors ($P \leq 0.20$) were Cancerstatus, Tobacco3, Differentiation, Recurrence, Age, Locoreg_fail and LN_positive. This forward selection has more factors than the backward elimination of Proc Phreg. The reasons can be the difference in P-value and the different methods used as before. Furthermore, in forward selection of Proc Lifetest, two hypothesis tests are preformed which are the log rank and Wilcoxon. Chem3 was significant for the log rank test but not Wilcoxon. As mentioned in the introduction, Wilcoxon is a weighted test that favors earlier events. Chem3, which is different types of chemotherapy, could be conducted in the later times. Since chemotherapy is typically given after primary surgery. In general, log rank test is far more popular.

When P16 status was treated as strata with other factors which include Age, Tobacco3, Recurrence, Locoreg_fail, Differnetiation, and LN_positve, the results were as expected.

For lymph nodes found positive (LN_positive), P16+ / less than 10 group had the best survival rate which was expected. Since P16+ group have better survival rate overall and less lymph node that has cancer cells, the better the survival for the patient.

The P16 and Age strata were also within expectation with P16+ group having better survival rate and younger groups have better survival as well. Tobacco status and P16 strata also were expected with P16+ having better survival and nonsmokers have better survival. With

Differentiation, many strata did not have many data point. However each stratum is significantly different with P16+ status as the best and survival rate for different stages of differentiation is typical with no differentiations, which are the earlier stages of caner having the best survival rate to poorly then moderately. In the group of Loco-regional fail, the strata seem to follow the survival trend of P16 status but do not show difference within loco-regional fail.

Since the dataset has relatively small number of patients, especially after censoring, several factor of interest were analyzed via strata in Proc Lifetest. Which were Chem3, Treatment5, Race, and Anemic. Treatment 5 was treated as strata over event time. It has 5 categories and maybe difficult to analyze with P16. Strata of treatment are significantly different with no treatment having the worst survival rate. The others are more or less similar in survival rate. However, types of treatment is still of interest since it is a factor that possess the most hope to patients. Therefore, chem3 was analyzed as strata. The category induction chemotherapy had the worst survival rate. Next race was analyzed and shoed difference with Caucasians having better survival rate.

How would types of treatment, types of chemotherapy and race differ in survival rate for the P16 groups? From the analysis, P16+ and P16− groups are similar in the strata. For treatment group, no treatment seemed to be worse in survival rate compared with the other treatment options, in P16+ groups, compared with P16− group, even the strata are significantly different, but not treatment group survival curve is much closer to the other groups. The interesting result is that Race is not significantly different in the subgroups when analyzed separately. This can be the result of not enough data in category 2 and 0.

Anemic was treated as strata as well. Result was as expected where P16+ groups had better survival rate with P 16+/ non-anemic group having the best survival rate. The interesting result is

that P16−/anemic group had better survival rate than P16−/none anemic group. However this is not significantly different on the 0.05 level and can be the result of small dataset.

For AFT models, gamma model followed by Weibull model had the least AIC, AICC, and BIC values which were expected since gamma model consists of the broadest assumption, and Weibull distribution model is derived from gamma model and the possesses the second broadest assumption. Exponential model assumes constant hazard which might not be the case for this study. L-normal is ideal for repeated events and this study is not designated for cancer repentance only. Last, L-logistic is best with binary data but this study has continuous variables as well as binary variables. In natural science, there are many unknown factors that should not make assumptions. Therefore, without may assumptions, gamma model should fit the dataset the best.

In conclusion, factors affected HNSCC (OPSCC) were as expected. Having treatment of any kind increase the chance of survival. In chemotherapy, induction chemo has the worst survival rate for this dataset. The early stages of HNSCC have better survival rate than the later stages. Recurrent cancer patients have worse survival rate. Younger patients have better survival rate as well. Similar to other studies, P16+ status had better survival rate and far better survival predictor than other factors.

References

1. Coordes A, Lenz K, Qian X, Lenarz M, Kaufmann AM, Albers AE. (2016) Meta-analysis of survival in patients with HNSCC discriminates risk depending on combined HPV and p16 status: Eur Arch Otorhinolaryngol 273: 2157–2169 DOI 10.1007/s00405-015-3728-0.
2. Baumeister P, Rauch J, Jacobi C, Kisser U, Betz C, BeckerS, Reiter M. (2016) Impact of comorbidity and anemia in patients with oropharyngeal cancer primarily treated with surgery in the human papillomavirus era: DOI 10.1002/hed.24528.
3. Stephen JK, Divine G, Chen KM, Chitale D, Havard S, and Worsham MJ. (2013) Significance of p16 in Site-specific HPV Positive and HPV Negative Head and Neck Squamous Cell Carcinoma (race): Cancer Clin Oncol, 2(1): 51–61. doi:10.5539/cco.v2n1p51.
4. Weinberger PM, Yu Z, Haffty BG, Kowalski D, Harigopal M, Sasaki C, Psyrri A. (2004) Prognostic significance of p16 protein levels in oropharyngeal squamous cell cancer. Clin Cancer Res. 10(17):5684–5691.
5. Lajer CB, Buchwald CV. (2010) The Role of Human Papillomavirus in Head and Neck Cancer. 010 APMIS DOI 10.1111/j.1600-0463.2010.02624.x.
6. D'Souza G, Dempsey A. (2011) The Role of HPV in Head and Neck Cancer and Review of the HPV Vaccine. rev Med. 53(Suppl 1): S5–S11. doi:10.1016/j.ypmed.2011.08.001.
7. Cutts FT, Franceschi S, Goldie S, Castellsague X, de Sanjose S, Garnett G, Edmunds WJ, Claeys P, Goldenthal KL, Harper DM, Markowitz L. (2007) Human papillomavirus and HPV vaccines: a review. Bulletin of the World Health Organization 85:719–726.

8. Kallogjeri D, Piccirillo JF, Spitznagel EL, Steyerberg EW. (2012) Comparison of Scoring Methods for ACE-27: Simpler Is Better. J Geriatr Oncol. 3(3): 238–245. doi:10.1016
9. Larsen C, Gyldenlove M, Jensen DH, Therkildsen MH, Kiss K, Norrild B, Konge L, Buchwald CV. (2013) Correlation between human papillomavirus and p16 overexpression in oropharyngeal tumours: a systematic review.
10. https://ghr.nlm.nih.gov/condition/head-and-neck-squamous-cell-carcinoma
11. http://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet
12. Kleinbaum DG, Klein M. (2013) Survival Analysis: A Self-Learning Text, Third Edition.
13. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M (2009): Defining Comorbidity: Implications for Understanding Health and Health Services. Annals of Family Medicine: 7(4):357-363. doi:10.1370/afm.983.
14. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_lifereg_sect019.htm
15. Sung W, Park JM, Choi CH, Ha SW, Ye SJ. (2012) The Effect of Photon Energy on Intensity-Modulated Radiation Therapy (IMRT) Plans for Prostate Cancer. Radiation Oncol J. 30(1): 27–35.
16. Strigaria L, Pinnaròb P, Carlinic P, Torinod F, Strolina S, Minossea S, Sanguinetib G, Benassie M. (2016) Efficacy and mucosal toxicity of concomitant chemo-radiotherapy in patients with locally-advanced squamous cell carcinoma of the head-and-neck in the light of a novel mathematical model. vCritical Reviews in Oncology/Hematology Volume 102, Pages 101–110.
17. http://www.cancer.org/treatment/treatmentsandsideeffects/treatmenttypes/chemotherapy/how-chemotherapy-drugs-work

18. http://wwwnc.cdc.gov/eid/article/16/11/10-0452_article

19. Allison PD. (2010) Survival Analysis Using SAS: A practical Guide 2nd Edition.

20. Hastie T, Tibshirani R, Friedman J. (2008) The Elements of Statistical Learning-Data Mining, Inference, and Prediction, 2nd Edition.

21. St Clair JM, Alani M, Wang MB, Srivatsan ES. (2016) Human papillomavirus in oropharyngeal cancer: The changing face of a disease: Biochimica et Biophysica Acta (BBA) - Reviews on Cancer**,** Volume 1866, Issue 2, Pages 141–150.

22. Okami K. (2016) Clinical features and treatment strategy for HPV-related oropharyngeal cancer: Int J Clin Oncol: DOI 10.1007/s10147-016-1009-6.

23. Sano D, Oridate N. (2016) The molecular mechanism of human papillomavirus-induced carcinogenesis in head and neck squamous cell carcinoma: Int J Clin Oncol: DOI 10.1007/s10147-016-1005-x.

24. Robinson M, Sloan P, Shaw R (2010) Refining the diagnosis of oropharyngeal squamous cell carcinoma using human papillo- mavirus testing. Oral Oncol 46:492–496