Statistical Analysis of Factors Associated with the Survival Time to Relapse in High-Risk Smokers

STAT 697 Issues in Statistics

Colin Busby

5/10/2023

1 ABSTRACT

Smoking is the leading cause of death in the United States, while smoking-related diseases afflict millions of Americans. Despite the dangers of smoking, it is challenging to quit. The University of Medicine and Dentistry of New Jersey performed a randomized clinical trial to evaluate the effectiveness of triple-combination pharmacotherapy for tobacco dependence treatment in smokers with medical illnesses who are susceptible to tobacco-related complications. This analysis aims to evaluate some factors that may affect a patient's risk of relapse during tobacco cessation. Both non-parametric methods like Kaplan-Meier Estimation and semi-parametric methods like the Cox Proportional Hazards model were used during this analysis and were performed using SAS 9.4. The analysis found that the most critical factors in a patient's risk of relapse were the treatment method, age, and employment status. The triple-combination therapy reduced a patient's risk of relapse by half compared to the patch-only treatment (hazard ratio = 0.544, p-value = 0.0053, 95% CI: (0.352, 0.832)). Additionally, every year older a patient was, their risk of relapse decreased by a factor of 0.965 compared to being a year younger (hazard ratio = 0.965, p-value = 0.0010). Finally, a patient's employment status had a significant impact on their risk of relapsing, with part-time employment causing an increase in risk by a factor of 1.926 (p-value = 0.0455) compared to full-time employment, while other types of employment increased the risk by a factor of 2.023 (p-value = 0.0089). Future research could analyze how stress may impact a patient's ability to quit smoking.

2 INTRODUCTION

2.1 BACKGROUND

According to the Centers for Disease Control and Prevention (CDC), smoking is the leading cause of death in the United States, with over 480,000 deaths a year. Additionally, over 16 million Americans are afflicted with a smoking-related disease. Smoking also exacerbates many illnesses, such as cancer and cardiovascular disease, leading to complications that may lead to death. Furthermore, smoking is tough to quit, with less than one in ten adults successfully quitting in 2018 despite 21.5 million smokers attempting to. A randomized controlled trial investigating the effectiveness of a triple-medication combination therapy versus the standard nicotine patch was carried out by the University of Medicine and Dentistry of New Jersey, Rutgers Cancer Institute of New Jersey, and the Robert Wood Johnson Foundation. The trial involved 127 smokers with predefined illnesses (including cardiovascular disease, cancer, diabetes, and others) from the local community, who were randomized into two equal groups, the triple-medication group and the nicotine patch-only group, for a 10-week-long tapering course. Patients were followed up at 26 weeks after the target quit date, with an abstinence rate of 35% (22 of 63 patients) from the combination group and 19% (12 of 64 patients) from the patch-only group. The median time to relapse was 65 days for the combination group and 23 days for the patch-only group.

2.2 DATA DESCRIPTION

The data set used in this analysis was collected from a randomized controlled trial by the

University of Medicine and Dentistry of New Jersey involving 127 smokers 18 years or older with predefined medical illnesses from the local community. Data was collected on eleven (11) prognostic variables (Table 1). The primary outcome measure was tobacco use abstinence, and secondary outcome measures were time to relapse, duration of medication use, and adverse clinical events. The time frame of the trial was 26 weeks. The endpoints of interest for this analysis are the survival time to relapse (in days) and relapse (0=censored and 1=relapse), with relapse indicating the patient resumed smoking. Before any analysis was carried out, all variables were recoded as numeric variables. During the investigation, the variables *ageGroup2* and *ageGroup4* were dropped due to being groupings of the variable *age*.

Variable	Description
id	Patient ID number
ttr	Time in days until relapse
relapse	Indicator of relapse (return to smoking)
grp	Randomly assigned treatment group with levels combination or patchOnly
age	Age in years at time of randomization
gender	Female or Male
race	black, hispanic, white, or other
employment	ft (full-time), pt (part-time), or other
yearsSmoking	Number of years the patient had been a smoker
levelSmoking	heavy or light
ageGroup2	Age group with levels 21-49 or 50+
ageGroup4	Age group with levels 21-34, 35-49, 50-64, or 65+
priorAttempts	The number of prior attempts to quit smoking
longestNoSmoke	The longest period of time, in days, that the patient has previously gone without smoking

Table 1 Description of Variables

2.3 RESEARCH QUESTION

This analysis aims to examine the association between various factors and the survival time to relapse for smoking patients undergoing tobacco dependence therapy, specifically a triple-medication combination versus standard-duration therapy using a nicotine patch.

2.4 STATISTICAL METHODS

For the data exploration section of this analysis, the non-parametric Kaplan-Meier Estimation was used first to compare the survival and hazard curves of the two treatment groups. Then, the semi-parametric Cox Proportional Hazard model was used to determine which covariates were statistically significant and their effect on the hazard risk**.**

2.5 MAIN OUTCOMES

Comparing the two treatment groups using the Kaplan-Meier estimation found a statistically significant difference between the triple-medication combination and the nicotine patch-only groups. Additionally, semi-parametric analysis using the Cox PH model found that the triple-medication combination decreased the risk of relapse by 54.4% (p-value < 0.005, hazard ratio = 0.544, 95% CI: (0.352, 0.832)) compared to the usage of only a nicotine patch.

3 STATISTICAL ANALYSIS

3.1 Kaplan Meier Estimate

To determine if there was a statistically significant difference between the two treatment groups, the non-parametric Kaplan Meier (K-M) estimate, stratified by the two treatment groups, was utilized with the results shown in Figure 1 below. The graph of the survival estimate curve shows that the survival probability of the combination treatment group (grpnum=1) is always greater than that of the patch group (grpnum=0), with minimal overlap of the confidence bands. Furthermore, as shown in Figure 3, the three test statistic p-values (0.0046, 0.0047, and 0.0003) are much smaller than the 0.05 threshold, indicating a statistically significant difference in the survival time between the two treatment groups. Additionally, Figure 2 shows the graphical check of the proportional hazards property, with the two curves being parallel without any crossing over, indicating that the proportional hazard assumption is satisfied.

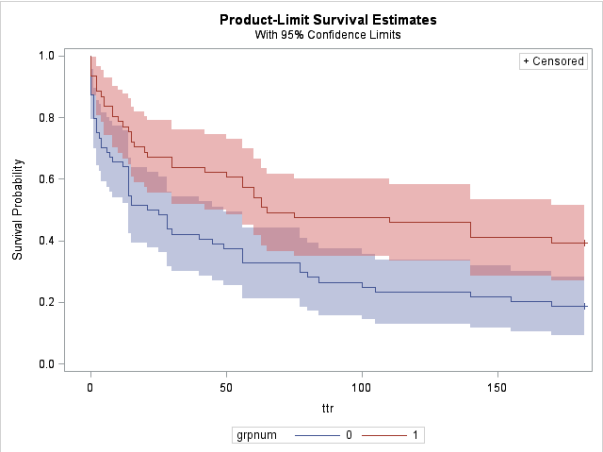

Figure 1 K-M Survival Estimates

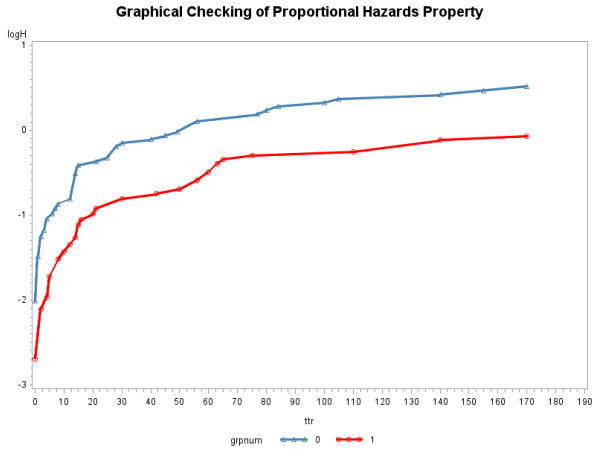

Figure 2 Proportional Hazards Checking

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	8.0276	1	0.0046
Wilcoxon	8.0097	1	0.0047
-2Log(LR)	13.2514	1	0.0003

Figure 3 Non-Parametric Hypothesis Tests

Hypothesis Test
H_0: $S_{trp}(t) = S_{patch}(t)$ vs. H_1: $S_{trp}(t) \neq S_{patch}(t)$, for all $t > 0$
Where t = time in days.
$S_{trp}(t)$: survival function of triple-combination group.
$S_{patch}(t)$: survival function of patch-only group.
Test Statistic and Decision: all p-values are < 0.05; therefore, H_0 is rejected
Conclusion: At alpha=0.05, there is a significant difference between the survival functions of the triple-combination and patch-only groups.

3.2 Model Selection

Initially, all covariates in the full model were tested under the non-parametric K-M estimation. The Wilcoxon and Log-Rank tests (both Univariate and Forward Stepwise Sequence) found that two (2) variables were consistently significant: *grpnum* and *age*. However, the variable *job* was found to be significant only in the Forward Stepwise tests. To corroborate the non-parametric results, the model selection was performed using the semi-parametric Cox Proportional Hazard model, utilizing backward selection and exact handling of ties. This model selection, shown in Figure 4, resulted in the selection of three variables: *grpnum* (p-value = 0.0053), *age* (p-value = 0.965), and *job* (p-value = 1.926 {job=1} and p-value = 2.023 {job=2}).
The final model's goodness-of-fit was tested by both the likelihood test and the AIC scores of the full and final models, as seen in Figure 5. The final model has an AIC of 632.169, which is smaller than the full model's AIC of 643.055, meaning that the reduced final model fits the data better than the full model. The likelihood test p-value of 0.7453 corroborates this, suggesting that there is no significant difference between the full and final models, indicating that the final model can be used instead of the full model.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
grpnum	1	1	-0.60854	0.21844	7.7611	0.0053	0.544	grpnum 1
age		1	-0.03536	0.01077	10.7802	0.0010	0.965	
job	1	1	0.65547	0.32773	4.0000	0.0455	1.926	job 1
job	2	1	0.70483	0.26954	6.8376	0.0089	2.023	job 2

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	Msmk	1	8	0.0451	0.8318
2	sex	1	7	0.0775	0.7807
3	priorAttempts	1	6	0.1490	0.6995
4	longestNoSmoke	1	5	0.4502	0.5022
5	racenum	3	4	3.5374	0.3159
6	yearsSmoking	1	3	0.4570	0.4990

Figure 4 Model Selection

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
-2 LOG L	646.223	624.169
AIC	646.223	632.169
SBC	646.223	642.123

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
-2 LOG L	646.223	619.055
AIC	646.223	643.055
SBC	646.223	672.919

Figure 5 Model Fit Stats: Final (left) and Full (right)

3.3 Assumption Check

The proportional hazards assumption was checked for violations using the PROC PHREG *assess ph/resample*. All p-values were non-significant (Figure 6), indicating that all covariates satisfied the proportional hazards assumption, consistent with the results in Section 3.1.

Supremum Test for Proportionals Hazards Assumption				
Variable	**Maximum Absolute Value**	**Replications**	**Seed**	**Pr > MaxAbsVal**
grpnum1	0.5008	1000	226778393	0.8490
job1	0.6518	1000	226778393	0.6040
job2	0.7020	1000	226778393	0.6890
age	0.8953	1000	226778393	0.4740

Figure 6 PH Assumption Test

3.4 Final Model and Interpretation

Using the Maximum Likelihood Estimates, the final model is:

$$\boldsymbol{h(y|x) = h_0(y) * exp\{-0.60854 * grpnum + 0.65547 * job1 + 0.70483 * job2 - 0.03536 * age\}}$$

where $\boldsymbol{h_0(y)}$ is the baseline hazard function and $\exp\{\beta_i x_i\}$ are the hazard ratios. According to the final model, when controlling for the other covariates, the effect of *grpnum* (0=patch only, 1=combination) has a relative risk of 0.544, indicating that a patient in the triple combination therapy group has a decreased risk of relapse by a factor of 0.544 compared to patients in the patch-only group. When controlling for the other covariates, the effect of *job* (0=fulltime, 1=part-time, 2=other) has a relative risk of 1.926 for part-time employment and a relative risk of 2.023 for other employment, indicating that a patient who has part-time work has an increased risk of relapse by a factor of 1.926, while a patient who has other employment has an increased risk of relapse by a factor of 2.023, when compared to a patient who has full-time employment. Finally, when controlling for the other covariates, the effect of *age* has a relative risk of 0.965, indicating that for a patient of age *a*, every one-year increment in age decreases the risk of relapse by a factor of 0.965 compared to the previous year.

4 CONCLUSION AND DISCUSSION

Based on the final model in Section 3.4, the covariates *grpnum*, *job*, and *age* all have a statistically significant effect on the hazard risk of relapse at the 0.05 significance level. Given that smoking is known to be challenging to give up, with patients often relapsing, the result that patients who undergo the triple combination therapy have their risk of relapse decreased by 54.4% when compared to only using a nicotine patch is a significant outcome. Further research into this topic could be beneficial for smokers trying to quit, especially those with preexisting illnesses found in the trial participants. Additionally, the finding that a person's employment status (such as full-time or part-time) has a massive effect on their chances of relapsing is quite important since this can be highly detrimental to a patient trying to quit for medical reasons. Further research in this area could focus on the effects of stress on a patient's chance of relapsing since part-time employment is often associated with a lower socioeconomic status, which is connected to higher stress levels. Finally, the impact of age on a patient's risk of relapsing should be studied further since the older a person gets, the more health complications appear, so an older patient may be more motivated to stop smoking, regardless of the method used.

5 APPENDICES

5.1 Appendix I

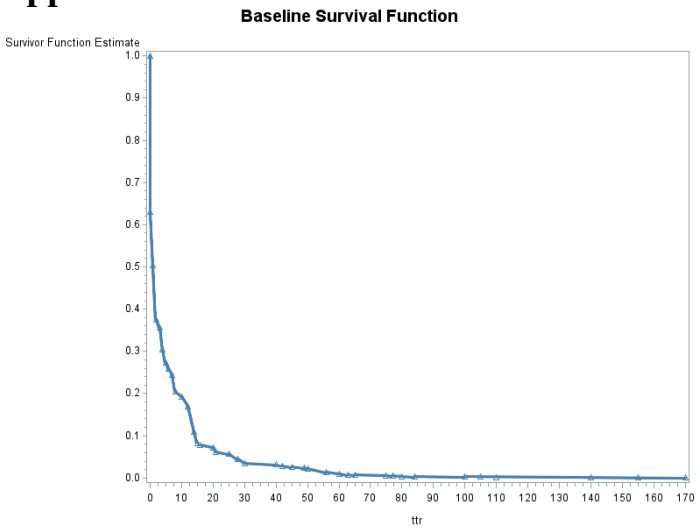

Figure 7 Baseline Survival Function

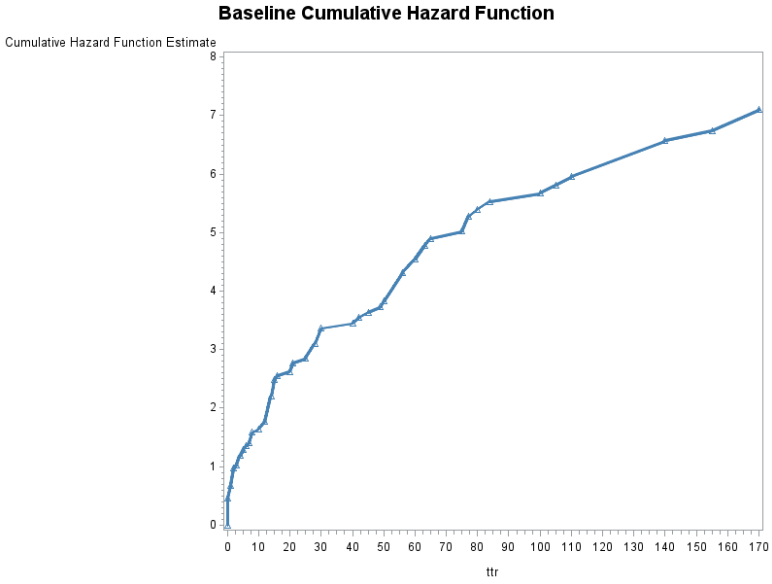

Figure 8 Baseline Cumulative Hazard Function

References

1. Dataset “pharmacoSmoking”: https://vincentarelbundock.github.io/Rdatasets/datasets.html
2. Centers for Disease Control and Prevention – Smoking & Tobacco Use
3. Steinberg, M.B. Greenhaus, S. Schmelzer, A.C. Bover, M.T., Foulds, J., Hoover, D.R., and Carson, J.L. (2009) Triple-combination pharmacotherapy for medically ill smokers: A randomized trial. Annals of Internal Medicine 150, 447-454.

SAS Code

```
proc import datafile="C:\Users\Home\Documents\My SAS
Files\9.4\STAT697_Final_Project\pharmacoSmoking.csv"
      out=smoke
      dbms=csv
      replace;
      getnames=yes;
run;

proc contents data=work.smoke;
run;

/* Recoding characters into numerics */
data smokenum;
      set smoke;
      if grp = "patchOnly" then grpnum = 0;
      if grp = "combination" then grpnum = 1;
      if gender = "Male" then sex = 0;
      if gender = "Female" then sex = 1;
      if race = "white" then racenum = 0;
      if race = "black" then racenum = 1;
      if race = "hispanic" then racenum = 2;
      if race = "other" then racenum = 3;
      if employment = "ft" then job = 0;
      if employment = "pt" then job = 1;
      if employment = "other" then job = 2;
      if levelSmoking = "heavy" then lvlsmk = 0;
      if levelSmoking = "light" then lvlsmk = 1;
      if ageGroup2 = "21-49" then ag2 = 0;
      if ageGroup2 = "50+" then ag2 = 1;
      if ageGroup4 = "21-34" then ag4 = 0;
      if ageGroup4 = "35-49" then ag4 = 1;
      if ageGroup4 = "50-64" then ag4 = 2;
      if ageGroup4 = "65+" then ag4 = 3;
run;

proc print data=smokenum;
run;

proc means data=smokenum median;
      class grp;
      var ttr;
run;

/* K-M estimation by strata using variable of interest grp*/
proc lifetest data=smokenum method=km conftype=linear
plots=(survival(cl),ls,lls)
      graphics outsurv=a;
      time ttr*relapse(0);
      strata grpnum;
run;

/* Checking covariates */
proc lifetest data=smokenum method=km conftype=linear
plots=(survival(cl),ls,lls)
```

```
      graphics;
      time ttr*relapse(0);
      test grpnum age sex racenum job yearsSmoking lvlsmk priorAttempts
longestNoSmoke;
run;

data a2;
      set a;
      s=survival;
      logH=log(-log(s));
      lnorm=probit(1-s);
      logit=log((1-s)/s);
      ldays=log(ttr);
run;

proc gplot data=a2;
      symbol1 i=join width=2 value=triangle c=steelblue;
      symbol2 i=join width=2 value=circle c=red;
      plot logit*ldays=grpnum logH*ldays=grpnum lnorm*ldays=grpnum;
run;

proc gplot data=a2;
      title1 "Graphical Checking of Proportional Hazards Property";
      plot logH*ttr=grpnum;
      symbol1 i=join width=2 value=triangle c=steelblue;
      symbol2 i=join width=2 value=circle c=red;
run;


/* Full Model */
proc phreg data=smokenum;
class job(ref='0') racenum(ref='0') grpnum(ref='0');
      model ttr*relapse(0) = grpnum age sex racenum job yearsSmoking lvlsmk
priorAttempts longestNoSmoke/ties=exact;
run;

/* Backwards model selection */
proc phreg data=smokenum;
class job(ref='0') racenum(ref='0') grpnum(ref='0');
      model ttr*relapse(0) = grpnum age sex racenum job yearsSmoking lvlsmk
priorAttempts longestNoSmoke/ties=exact selection=backward;
run;

/* Checking PH assumption of fitted final model */
proc phreg data=smokenum;
class job(ref='0') grpnum(ref='0');
      model ttr*relapse(0) = grpnum job age/ties=exact;
      hazardratio grpnum/diff=ref CL=both;
      assess ph/resample;
run;


/* Baseline survival function */
data null;
      input grpnum job age;
      cards;
      0 0 0
```

```
run;

proc phreg data=smokenum;
class job(ref='0') grpnum(ref='0');
      model ttr*relapse(0) = grpnum job age/ties=exact;
      baseline out=b covariates=null survival=s lower=lcl upper=ucl cumhaz=H
lowercumhaz=lH uppercumhaz=uH;
run;

proc print data=b;
run;

/* Baseline survival & cumulative hazard functions */
proc gplot data=b;
title "Baseline Survival Function";
plot s*ttr;
run;
proc gplot data=b;
title "Baseline Cumulative Hazard Function";
plot H*ttr;
run;

/* Calculating p-value for goodness of fit */
data chi;
      y=1-cdf('CHISQ', 5.114, 8);
      put 'p-value:  ' y;
run;

proc print data=chi;
run;

quit;
```