Timing of Adjuvant Chemotherapy and Survival Outcome After Lung Cancer Surgery

Issac Li, Final Project for Categorical Data Analysis

Methods

Data Source

The dataset used in this study is from the National Cancer Data Base on patients that received adjuvant chemotherapy after non-small cell lung cancer (NSCLC) resection.

Sample Population

In order to study the survival after resection with adjuvant chemotherapy with generalized linear regression models, an outcome variable indicating death at 4 years was generated and assigned based on the patient's survival months (numeric) and vital status (binary 0, 1). Patients who have survival time less than or equal to 48 months and vital status 0 (dead) are assigned 0 for the outcome variable. These patients are the cases. Patients who have survival time greater than 48 months are assigned 1 for the outcome variable. These patients are the controls. Total population of the dataset is 12473. Number of patients in the case group (with outcome 1) is 4396 and number of patients in the control group is 4318. The rest 3759 samples (patients who had resection within 4 years and were still alive when the dataset was accessed) are removed from the total population to form a sample population of 8714. All the sample population have same primary type of treatment and radiation so these two variables are excluded for this study.

Data Elements

The NCDB dataset contained variables that gave comprehensive information about each patient. These variables include indicators of physiological status of the patients as well as non-physiological information of them. Physiological independent variables included were: age, sex, Charlson-Deyo (CD) score (0, 1, ≥2), year of diagnosis, tumor primary site, histology, histological grade of tumor, tumor size, pathological stage, resection type (lobectomy or pneumonectomy), days until start of post-operative chemotherapy after resection, vital status of the patients, and number of Months after resection when vital status was recorded. Variables covering non-physiological information included facility type (academic, non-academic) and location (geographic region that the patient's treating facility was located in), insurance status, income (median income of the patient's area of residence), education (percent of people in the patient's area of residence with no high-school diploma) and year of diagnosis. The number of days before adjuvant chemotherapy is considered physiologically relevant because this time is usually related to the recovery of patients after resection operation.

Statistical Analysis

Imputation of Missing Data

There are 3.64% of the grades recorded as unknown. Since histological grade has been reported to be predictive of cancer survival outcome, [1] it may be beneficial to first impute the unknown grades. The unknown grades should not be regarded as a class of grades for classification purposes because the causes of grades to be unknown are usually insufficient biopsy samples or loss of data. So theoretically the unknown grades would become known if we had supplied the missing data. Thus ordered logistic regression was performed using library MASS in R version 3.3.1 (R Foundation).

Ordinary Logistic Regression with Grouped Samples

As is shown in the result section, there is obvious clustering due to some of the non-physiological variables. Thus logistic regression on the whole sample data would be influenced by confounding effects. To get clearer picture of the role of the physiological variables in the survival of patients, we subset the sample data into two subgroups and then used the glm() function in R to perform ordinal logistic regression with one shared formula modified from the formula given by backward selection.

Mixed Effects Logistic Regression

Another way to address the problem of clustering is to use generalized mixed effects modeling (GLMM) which accounts for both fixed effects and random effects. In this case, we consider the physiological variables to have fixed effects since values of those variables are inherent to each patient. On the other hand, we consider the random effects to be the non-physiological variables, since the hospitals and doctors may have different experience, expertise and toolsets for the resection and radiation treatment. In this dataset, this is no doctor-level data, but there is information on the types of hospital that the patients go to, such as facility type and location. We chose to model these two variables as nested random effects. The mixed effects modeling is implemented using the "lmer" library in R.

For all types of regression models, inference is calculated in R. For the mixed effect and ordinary models, odds ratios, 95% confidence intervals and predictive powers are calculated. For the ordered logit models and ordinary logistic models, χ2 tests and goodness-of-fit tests are performed.

Results

Sample Characteristics

To find the optimal interval, we grouped days before chemotherapy into weeks (designated `Chemo_weeks` in model). The number of weeks range from 3 to 19. But for simplicity and balance of observations in each category **(Figure 1B)**, we merged week 3 and 4 and merged every week after 9 into week 9, so week 9 actually means week 9 and after. Exploratory graphical analysis **(Figure 1C)** and simple logistic regression (results not shown) were not able to give useful information about the role of time interval before chemotherapy in survival outcome. Other graphical analysis yielded important insights on the effects of non-physiological factors on survival outcome (**Figure 2**). These observations led us to

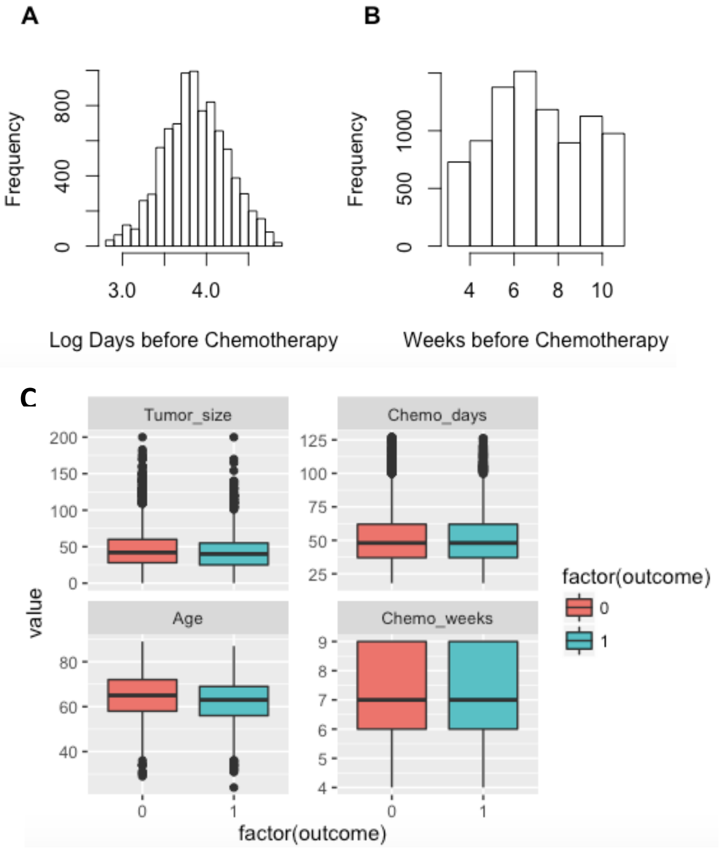

Figure 1A Distribution of Days before Chemotherapy. The logged days are normally distributed. **1B Distribution of Weeks before Chemotherapy.** Numbers of patients in each week are not too different. **1C Graphical Exploratory Analysis of Relationship between Continuous Variables and Outcome.** Age and tumor size look negatively correlated to outcome while chemo days and weeks seem uncorrelated to outcome.

treat clustering by 1) subgrouping the dataset by some non-physiological factors and 2) using GLMMS. Characteristics of subgroups defined by the optimal interval given by regression analysis are summarized in **Table 1.**

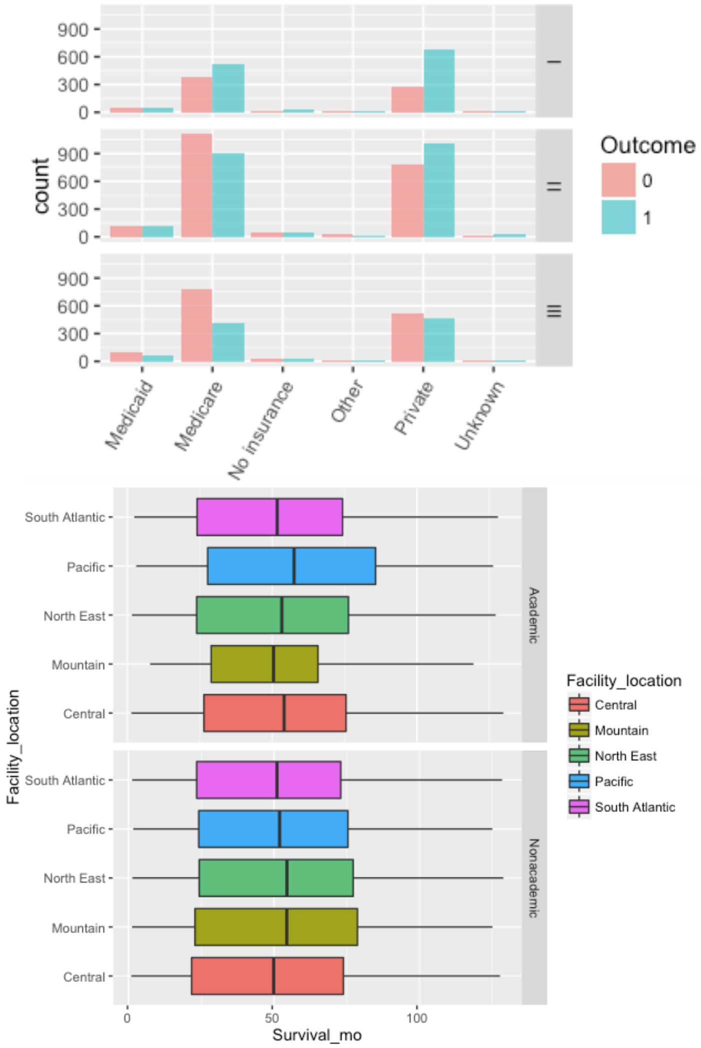

Figure 2 Upper Panel. Distribution of Outcomes for Patients with Different Insurance Providers. We see that patients with private insurance on average have better outcome than patients with Medicare in all three stages of NSCLC. **Bottom Panel. Boxplot Showing Median Survival Month for Patients Treated in Different Hospitals.** Different types of hospital in different region perform differently. Notice the difference between academic hospitals and non-academic ones in central region and greater variance within academic hospitals.

Imputation of Missing Grade

There are 408 patients with grade 1 tumor, 3534 with grade 2, 4198 with grade 3 and 257 with grade 4. Unknown grades are 317 which is more that grade 4 cancer and close to grade 1. Imputation of unknown grades was achieved by ordered logistic regression with the formula:

$$\mathit{Grade} \sim \mathit{Tumor\_size} + \mathit{Histology} + \mathit{Primary\_site} + \mathit{Surgery} + \mathit{Sex}$$

This formula was selected from a full model up to all two-way interaction terms by backward elimination based on AIC. Training dataset included randomly selected 75% of all the samples whose grades are known. Testing dataset was the remaining 25% whose grades are known. Prediction accuracy within the training dataset was 54.28% and 53.04% within the testing dataset. Out of the 317 unknown grades, 110 were imputed as grade 2 and 207 were imputed as grade 3. None was imputed as grade 1 or 2.

Logistic Modeling of Grouped Sample

Ordinary logistic regressions on the whole dataset with imputed grades were fruitless in giving significant intervals (results not shown). So we hypothesized that there are confounding factors that group the whole dataset into smaller, more homogenous ones. We performed a logistic regression of the socioeconomic variables onto the outcomes to get the significance of each variables quantitatively. Variables whose P-value were below or about 0.05 were education (0.054), insurance (<< 0.001), and facility type. We roughly divided the dataset into two: hospital type being academic (model 2) versus hospital type being non-academic (model 1). The rationale for this is that we observed significant discrepancy (p = 0.0029) on the survival months between patients treated in these two types of hospitals. Also, literature has reported difference in performance within the two. [2] Both subgroups were fitted with the same formula:

$$\mathit{Outcome} \sim \mathit{Age} + \mathit{Sex} + \mathit{CD\_score} + \mathit{Path\_stage} + \mathit{Grade} + \mathit{Tumor\_size} + \mathit{Surgery} + \mathit{Chemo\_weeks} + \mathit{Histology} + \mathit{Primary\_site} + \mathit{Age{:}CD\_score} + \mathit{Sex{:}CD\_score} + \mathit{Sex{:}Surgery} + \mathit{CD\_score{:}Path\_stage}$$

Table 1. Patient Characteristics of the Different Time-Interval Groups. Unknown entries in Income and Education are omitted for simplicity.

Characteristics	Early Start	Optimal	Delayed Start	P Value
	N=3132 (35.9%)	N=1515 (17.4%)	N=4067 (46.7%)	
Age (mean (sd))	62.67 (9.51)	63.62 (9.19)	63.99 (9.40)	<0.001
Survival Month (mean (sd))	49.57 (31.82)	50.06 (31.74)	47.50 (31.15)	0.004
Tumor Size (mean (sd))	44.40 (30.60)	45.40 (25.75)	46.47 (31.62)	0.016
Chemo Days (mean (sd))	33.41 (6.61)	45.44 (4.31)	67.43 (16.61)	<0.001
Outcome = 1 (%)	1590 (50.8)	795 (52.5)	2011 (49.4)	0.120
Sex = Male (%)	1740 (55.6)	788 (52.0)	2145 (52.7)	0.023
CD Score (%)				0.043
0	1708 (54.5)	851 (56.2)	2117 (52.1)	
1	1075 (34.3)	489 (32.3)	1464 (36.0)	
2+	349 (11.1)	175 (11.6)	486 (11.9)	
Pathological Stage (%)				<0.001
I	684 (21.8)	331 (21.8)	1021 (25.1)	
II	1575 (50.3)	774 (51.1)	1875 (46.1)	
III	873 (27.9)	410 (27.1)	1171 (28.8)	
Chemo Weeks (%)				<0.001
1	729 (23.3)	0 (0.0)	0 (0.0)	
2	988 (31.5)	0 (0.0)	0 (0.0)	
3	1415 (45.2)	0 (0.0)	0 (0.0)	
4	0 (0.0)	1515 (100.0)	0 (0.0)	
5	0 (0.0)	0 (0.0)	1189 (29.2)	
6	0 (0.0)	0 (0.0)	2878 (70.8)	
Grade (%)				0.124
1	157 (5.0)	71 (4.7)	180 (4.4)	
2	1272 (40.6)	647 (42.7)	1713 (42.1)	
3	1606 (51.3)	740 (48.8)	2071 (50.9)	
4	97 (3.1)	57 (3.8)	103 (2.5)	
Histology (%)				<0.001
Adenocarcinoma	1795 (57.3)	814 (53.7)	2137 (52.5)	
Large Cell Carcinoma	170 (5.4)	79 (5.2)	198 (4.9)	
Other	216 (6.9)	99 (6.5)	246 (6.0)	
Squamous Cell Carcinoma	951 (30.4)	523 (34.5)	1486 (36.5)	
Primary Site (%)				0.065
Lower Lobe	1082 (34.5)	483 (31.9)	1389 (34.2)	
Lung, NOS	57 (1.8)	18 (1.2)	78 (1.9)	
Middle Lobe	132 (4.2)	68 (4.5)	179 (4.4)	
Overlapping Lesion	73 (2.3)	42 (2.8)	136 (3.3)	
Upper Lobe	1788 (57.1)	904 (59.7)	2285 (56.2)	
Surgery = Pneumonectomy (%)	387 (12.4)	209 (13.8)	612 (15.0)	0.005
Facility Type = Nonacademic (%)	2252 (71.9)	1011 (66.7)	2530 (62.2)	<0.001
Facility Location (%)				0.003
Central	1522 (48.6)	707 (46.7)	1873 (46.1)	
Mountain	114 (3.6)	42 (2.8)	102 (2.5)	
North East	519 (16.6)	295 (19.5)	819 (20.1)	
Pacific	219 (7.0)	108 (7.1)	299 (7.4)	
South Atlantic	758 (24.2)	363 (24.0)	974 (23.9)	
Insurance (%)				<0.001
Medicaid	151 (4.8)	74 (4.9)	252 (6.2)	
Medicare	1372 (43.8)	714 (47.1)	2018 (49.6)	
No insurance	64 (2.0)	34 (2.2)	119 (2.9)	
Other	35 (1.1)	14 (0.9)	29 (0.7)	
Private	1466 (46.8)	662 (43.7)	1608 (39.5)	
Income (%)				0.448
<$48K	1411 (45.1)	646 (42.6)	1817 (44.7)	
$48K+	1668 (53.3)	842 (55.6)	2168 (53.3)	
Education (%)				0.412
<7%	647 (20.7)	355 (23.4)	831 (20.4)	
13-20.9%	901 (28.8)	409 (27.0)	1136 (27.9)	
21%+	469 (15.0)	226 (14.9)	637 (15.7)	
7-12.9%	1063 (33.9)	499 (32.9)	1384 (34.0)	

This formula was again reduced from a full formula with all terms up to all two-way interaction terms with all the physiological variables. None of the non-physiological variable was included between the subgroups were grouped by facility type and also we believe that non-physiological variables should not be considered by doctors when making treatment strategies (see more in discussion section). Hosmer and Lemeshow goodness-of-fit test on model 1 yielded $\chi2$ = 4.89, DF = 8, p-value = 0.5189; on model 2 yielded $\chi2$ = 10.14, DF = 8, p-value = 0.6222. Since both of the p-values >> 0.05, there was no evidence showing lack of fit. Prediction accuracy of model 1 was 60.57% at threshold = 0.5. The corresponding full model of model 1 predicted 63.1% of the outcomes accurate but a Chi-square test of the two models yielded 252.21 reduced deviance on 256 *df*, which is not significant (p-value = 0.216). Thus we still preferred the simpler model 1. The same applied to model 2. Prediction accuracy of model 2 was 63.68%. ROC curves of both models are shown **(Figure 3).**

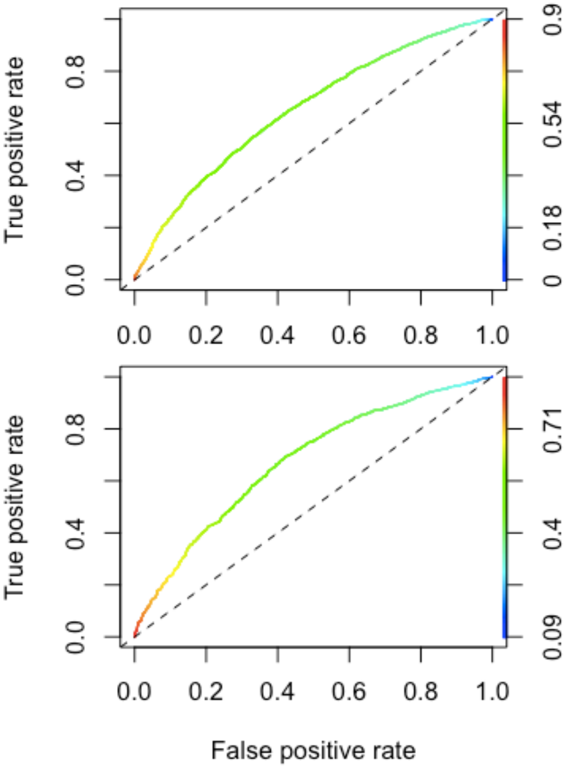

Figure 3 Upper Panel. ROC Curve of Model 1. Bottom Panel. ROC Curve of Model 2. Based on the ROC curves, model 2 is fitted its data better than model 1. The less zic-zac pattern of curve 1 means that model 1 might need more higher-order or interaction terms.

Model 1 estimated the most significant (**Table2,** p-value close to 0.1) decrease in relative survival odds in week 8 (50-56 chemo days) and the 95% CI of week 7 (43-49 days) gives the highest range of odds ratio to reference week 4 (less than or equal to 28 days). Model 2 estimated the most significant increase in odds of survival in week 7 (p-value < 0.05) with 95% CI being (1.103, 2.267). Week 5 has an even higher odds ratio, but the only reconciled observation across the two models is high odds ratio for survival during week 7.

Mixed Effects Model

As is evident from the two logistic models and exploratory analysis, the dataset is indeed nested by some random effects not inherent to each patient. These random effects are accounted for in GLMMs. The formula used to fit the GLMM was:

$$Outcome \sim Path\_stage + scaled\ Tumor\_size + Age + Chemo\_weeks + Histology + Grade + (1 \mid Facility\_location) + (1 \mid Year\_diag) + (1 \mid Facility\_type) + (1 \mid Income)$$

The last four terms indicate the random effect terms and all the other terms are fixed effect terms. Year diagnose contributed the largest and abnormal variance of nearly 10 when all data were included. Its variance was reduced to 0.06 ± 0.24 when data after 2010 were excluded (See more in discussion). The estimates for fixed effects were not impacted significantly by the inclusion/exclusion of these data, which would be impossible in ordinary regression models. The GLMM model outputs, including calculated p-values and rough estimates of CIs, are summarized in **Table 2**.

Odds ratios between Week 7 and Week 4 is the highest among all other weeks and the only one statistically significant at the p = 0.1 level. This means that the odds of survival at 4 years if adjuvant chemotherapy is started 4 weeks (or 49-56 days) after the NSCLC resection are 1.19 times (95% CI, 0.99, 1.45) the odds of survival if after 1 week. Prediction accuracy at *P = 0.5* is 62.21%. Plot of prediction distribution is also shown (**Figure 4**). Year of diagnose, insurance type and facility type did contribute much variance into the data. Also play roles in affecting the expected outcomes. Predicted probabilities of survival for different patients over different starting weeks are shown in **Figure 5**.

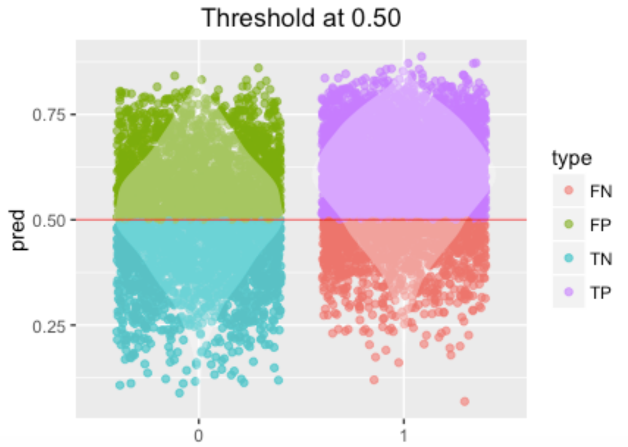

Figure 4 Prediction Type distribution.

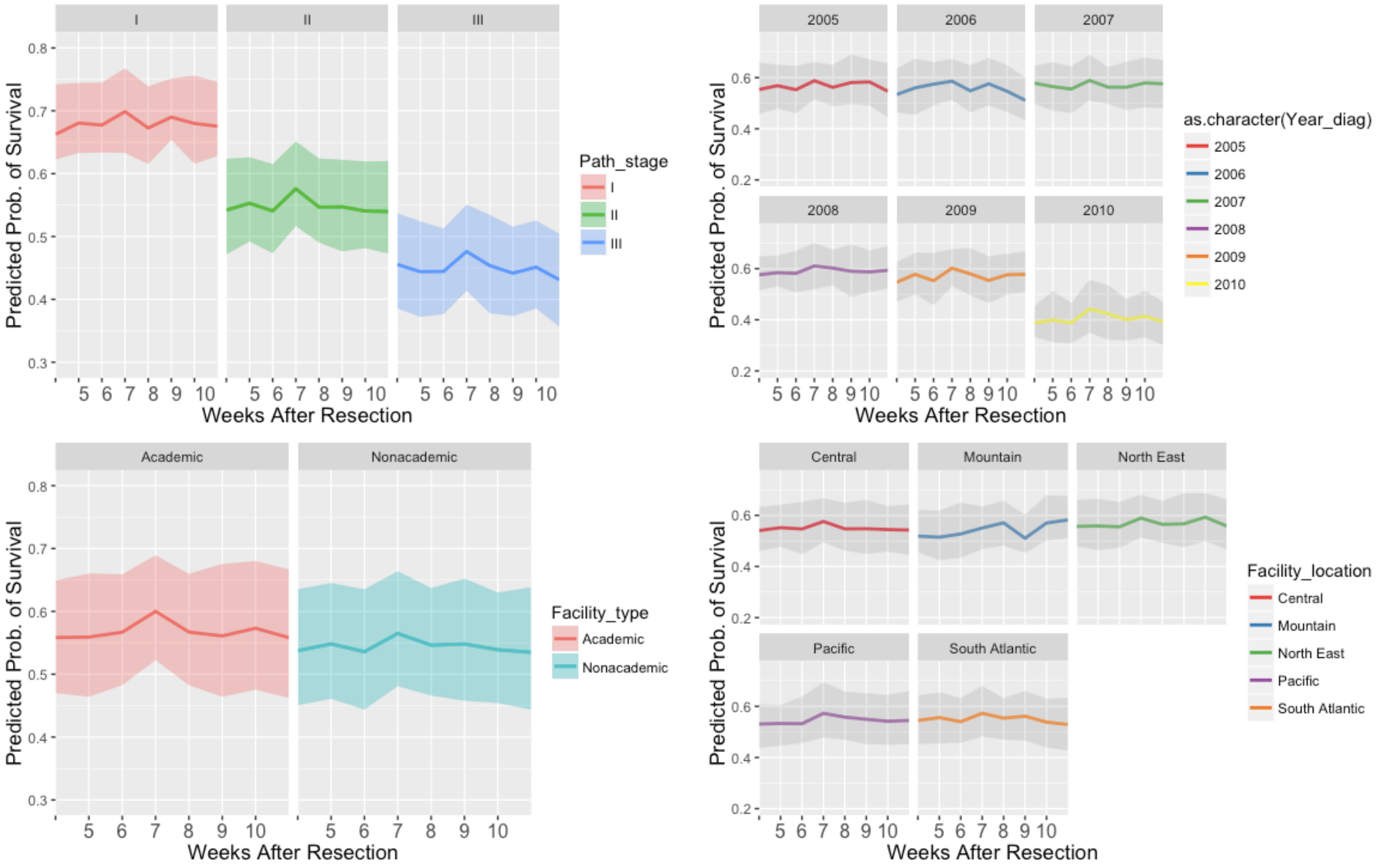

Figure 5. Predicted Probabilities of Survival with 90 % Interval. Upper-Left Panel. Cancer patients with all three stages of cancer have highest survival probability with optimal time point 49 days (interval 42 to 56) days between resection and adjuvant chemotherapy. Higher cancer stage correlates with lower outcomes. **Upper-Right Panel.** Cancer patients diagnosed across 2005 to 2010 all have higher survival probabilities in the interval 42 to 56 days with the optimal point 49 days. Variance across years are high. **Lower-Left Panel.** Academic hospitals treat patients with better outcomes. Patients treated at both academic and non-academic hospitals share similar optimal time interval. **Lower-Right Panel.** Patients treated at hospitals in different geographic regions slightly different outcomes and roughly similar optimal intervals.

Table 2. Model Outputs and Inference from the Two Logistic Regression Models on Subgroups. Bolded rows are of interest. * denotes significance at 0.01 level, ** denotes significance at the 0.05 level.

		ESTIMATE	STD. ERROR	Z-VALUE	PR(>|Z|)	ODDS RATIOS	2.5%	97.5%
MODEL 1	Week 5	-0.072	0.120	-0.602	0.547	0.930	0.736	1.176
	Week 6	-0.070	0.110	-0.638	0.524	0.932	0.751	1.157
	Week 7	**0.012**	**0.110**	**0.112**	**0.911**	**1.012**	**0.816**	**1.256**
	Week 8	**-0.165**	**0.115**	**-1.430**	**0.153**	**0.848**	**0.677**	**1.063**
	Week 9	-0.111	0.127	-0.874	0.382	0.895	0.697	1.148
	Week 10	-0.101	0.119	-0.852	0.394	0.904	0.716	1.140
	Week 11	-0.025	0.123	-0.201	0.841	0.976	0.767	1.242
MODEL 2	**Week 5**	**0.579**	**0.207**	**2.800**	**** 0.005**	**1.785**	**1.190**	**2.678**
	Week 6	0.227	0.189	1.201	0.230	1.255	0.866	1.818
	Week 7	**0.458**	**0.184**	**2.495**	*** 0.013**	**1.582**	**1.103**	**2.267**
	Week 8	0.275	0.190	1.452	0.147	1.317	0.908	1.909
	Week 9	0.346	0.193	1.798	0.072	1.414	0.969	2.062
	Week 10	0.299	0.189	1.585	0.113	1.349	0.932	1.952
	Week 11	0.159	0.192	0.832	0.405	1.173	0.806	1.707
GLMM	Week 5	0.032	0.107	0.295	0.768	1.032	0.836	1.274
	Week 6	0.033	0.100	0.332	0.740	1.034	0.850	1.257
	Week 7	**0.177**	**0.099**	**1.802**	*** 0.072**	**1.194**	**0.985**	**1.449**
	Week 8	0.050	0.103	0.488	0.626	1.052	0.859	1.288
	Week 9	0.065	0.110	0.596	0.551	1.068	0.861	1.324
	Week 10	0.079	0.105	0.756	0.449	1.082	0.882	1.329
	Week 11	0.029	0.108	0.274	0.784	1.030	0.834	1.272

Discussion

In this study, we are able to use generalized linear models to identify the time interval between NSCLC resection and adjuvant chemotherapy when the treatment effect (survival outcome) is optimal. When fitting the regression models, none of the socioeconomic variables, insurance provider, average income and education of resident area are used to fit the models. The reason of exclusion is logical rather than statistical: although these socioeconomic variables have significant effects on the outcome success, it would be unreasonable and controversial for doctors to adjust treatment regimens based on patients' socioeconomic status when cost is not a differential barrier, as the time of imitating the treatment does not change the cost of the treatment.

The GLMM on the whole sample space yields a statistically significant optimal time interval at the p = 0.1 level, bounded by the two significant intervals predicted by the two ordinal logistic models on the subgroups. This observation means there are nested samples in the dataset caused by random effects and ordinary logistic regression is inadequate for samples like this. However, the disadvantage of GLMM lies in its difficulty to interpret, Inference from GLMMs is complicated. When dividing up the samples into multiple levels, observations at some levels may be insufficient to guarantee the assumption that estimate over its standard error (SE) is normally distributed. Other methods can be used for inference for GLMMs include Monte Carlo simulation, Bayesian estimation, and bootstrapping, but these can be complex to implement. In this study, we did not use any of this method.

Instead we only estimated the CIs roughly with the SEs. This is one limitation of this study.

Another limitation of the study is that when engineering the outcome variable, we essentially excluded patients whose vital status was 1 and who were diagnosed after 2010. Only patients whose vital status was 0 from the same period were included. This asymmetry may be problematic since we have found that for some unknown reason, the year of diagnose plays a statistically significant role here. By excluding patients from 2011 and 2012 from control samples, we introduced significantly greater variance to the samples by year of diagnose and this would change the regression models if we do not also remove case samples from the same period. Thirdly, the time intervals are treated as factor variables in this study, thus inference and CIs on the optimal interval is impossible. And finally, the assumption of linearity for logistic regression is questionable in this case since some predictor variables are correlated **(Figure 6)**. For example, people with larger tumor sizes tend to have pneumonectomy over lobectomy, and pneumonectomy usually requires longer time to recovery thus delaying the time for start of chemotherapy. Because of these limitations, other regression models capable of finding threshold in continuous risk function which does not assume linearity should be further investigated to find a “real” interval for optimal treatment effect, such as Cox model with polynomial terms, which has been used to model treatment for colon cancers. [3]

Reference

1. Sun Z, Aubry M-C, Deschamps C, et al. Histologic grade is an independent prognostic factor for survival in non-small cell lung cancer: an analysis of 5018 hospital- and 712 population-based cases. *J Thorac Cardiovasc Surg*. 2006;131(5):1014-1020. doi:10.1016/j.jtcvs.2005.12.057.

2. Contemporary Performance of U.S. Teaching and Nonteaching Ho... : Academic Medicine. LWW. http://journals.lww.com/academicmedicine/Fulltext/2012/06000/Contemporary_Performance_of_U_S__Teaching_and.13.aspx. Accessed January 16, 2017.

3. Sun Z, Adam MA, Kim J, et al. Determining the Optimal Timing for Initiation of Adjuvant Chemotherapy After Resection for Stage II and III Colon Cancer. *Dis Colon Rectum*. 2016;59(2):87-93. doi:10.1097/DCR.0000000000000518.