BIOS 780 Final Project: Survival Analysis of Colon Cancer study

Chanhwa Lee, Dept. of Biostatistics, University of North Carolina at Chapel Hill

December 7, 2021

1 Introduction

More than 100,000 Americans are afflicted by colon cancer every year, and 80% of early stage cancer can be removed surgically. However, half of patients who have regional nodal involvement of the disease (referred as Stage C disease / colorectal cancer) might have recurrence of the disease within 5 years even after getting clinical resection. Until 1990, the best standard therapy was surgery alone, but some reports (*J. Clin. Oncology* 1989;7:1447-1456 and *New England J. of Medicine* 1990;322:352-358) suggested that the use of levamisole in combination with 5-fluorouracil (5-FU) might be beneficial for Stage C colorectal cancer.

In this report, we have analyzed the data on the Stage C patients who were reported in the *J. Clin. Oncology* and in the *New England Journal of Medicine* article to address the influence of treatment on time to disease recurrence and on patient survival (time to death). Furthermore, we have discussed the natural history of this disease based on a large and rich resource of the dataset, building a parsimonious model which can be used to predict survival of future patients. We also present the methods for model validation, in terms of the prediction of our model to new dataset.

The data is comprised of N = 1172 subjects with treatment assignment, survival status at the end of the study, recurrence status of the disease, various date (study registration, treatment assignment date, etc.), 12 clinical covariates, and data source. One part of data (929 subjects) is from SWOG 8591 (aka Int - 0035) and the other part (243 subjects) is from NCCTG. Missing is rare in the first part of the data (except for pre-operative CEA level and Nodal involvement covariates), but in the second part, 6 covariates (Location of primary neoplasm, Histologic type, Differentiation, Extent of local spread, Regional implants, pre-operative CEA level) are totally missing and a few covariates are partly missing. Therefore, we decided to use the first part of the data as model building set (namely, group 1), and the other part of the data as model validation set (namely, group 2).

We first show that the combination of levamisole and 5-fluorouracil has a positive effect on prognosis as well as patient survival, then use the data from the model building set to develop a natural history model. Such a model will be useful not only in counseling patients and in understanding the course of colon cancer in untreated patients, but also in providing historical control information to evaluate the efficacy of new therapeutic interventions. The data set of 243 patients from group 2 is used in model validation of the natural history model and to illustrate its use in survival prediction.

2 Influence of treatment

2.1 Treatment effect on time to disease recurrence

Figure 1 presents the Kaplan-Meier estimates of disease recurrence of colon cancer patients by treatment. The curves for observation and Levamisole group show little separation, while Lavamisole with 5-fluorouracil group has a distinct curve. Under the proportional hazards assumption, the Cox regression model can be used to measure treatment effect. If treatment is coded by $(Z_1, Z_2) = (1, 0)$: Levamisole, $(Z_1, Z_2) = (0, 1)$: 5-FU + Levamisole, $(Z_1, Z_2) = (0, 0)$: Observation, then in the model

$$\lambda(t|Z_1, Z_2) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2),$$

$\lambda_0(t)$ represents the hazard function for disease recurrence while not treated (observation), and β_1 is the log of the hazard ratio of Levamisole treated group; i.e., if $\lambda_1(t) \equiv \lambda(t|(Z_1, Z_2) = (1, 0))$ then, for all t,

$$\lambda_1(t)/\lambda_0(t) = e^{\beta_1}.$$

Similarly, β_2 is the log of the hazard ratio of 5-FU + Levamisole treated group; i.e., if $\lambda_2(t) \equiv \lambda(t|(Z_1, Z_2) = (0, 1))$ then, for all t,

$$\lambda_2(t)/\lambda_0(t) = e^{\beta_2}.$$

The maximum partial likelihood estimate for β_1 is $\hat{\beta}_1 = -0.01516$ (se: 0.10708, p-value = 0.887) and that for β_2 is $\hat{\beta}_2 = -0.51191$ (se: 0.11863, p-value = 1.59e-05). Under the proportional hazards assumption, a 95% confidence interval for hazard ratio of Levamisole group is (0.7985, 1.2150) and that of 5-FU + Levamisole group is (0.4750, 0.7562), indicating that the disease recurrence rate on Levamisole group to be 0.9850 that on observation group, but not statistically significant. However, we estimate the disease recurrence rate on 5-FU + Levamisole group to be 0.5993 that on observation group with a statistical significance. The proportional hazards assumption looks reasonable in this case because the Kaplan-Meier curves of 5-FU + Levamisole group and Observation group do not cross. Even though the curves of Levamisole group and Observation group cross each other, it is because there is no statistical evidence that these two curves are distinguishable. Therefore, we conclude that the proportional hazards assumption holds for this model, and indeed it is well fitted.

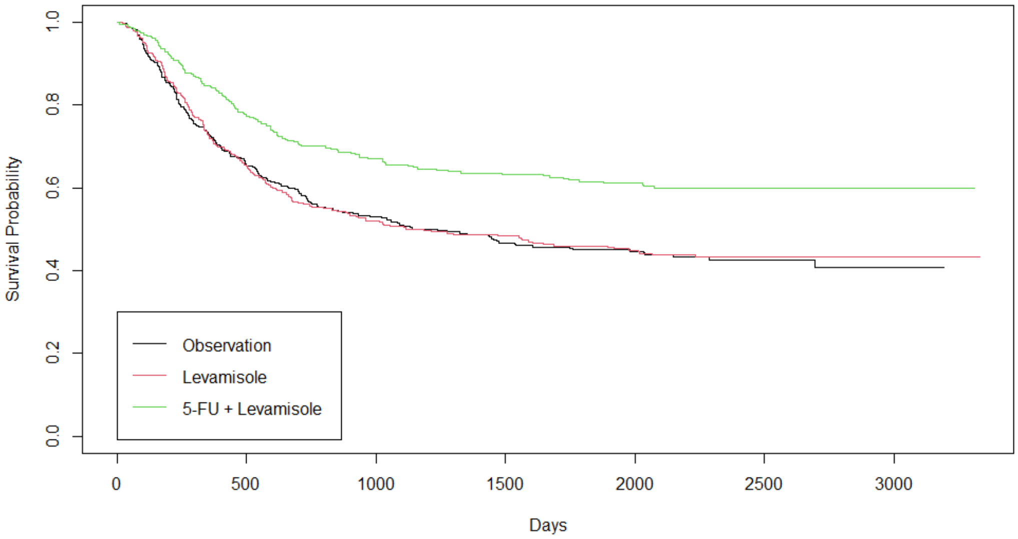

Figure 1: Kaplan-Meier curves of disease recurrence by treatment groups

We also conducted log rank tests for comparing treatment effects. For comparing Levamisole versus observation, patients in Levamisole and observation groups were used to test the difference between Kaplan-Meier curves (i.e., patients in 5-FU + Levamisole group are not used) using log rank test. Comparing 5-FU + Levamisole versus observation and 5-FU + Levamisole versus Levamisole only were also conducted similarly. Table 1 presents test results, indicating that there is no difference between KM curves of Levamisole and observation group, but there is difference between that of 5-FU + Levamisole group and other groups. Thus, we can conclude that there is evidence of 5-FU + Levamisole has positive effect on prognosis.

Table 1: Log rank tests for comparing treatment effects on disease recurrence

Comparsion	Chisq test statistic	DF	P-value
Lev vs. Obs	0.022	1	0.9
5-FU + Lev vs. Obs	19.1	1	1e-05
5-FU + Lev vs. Lev	17.7	1	3e-05

2.2 Treatment effect on time to death

Figure 2 presents the Kaplan-Meier estimates of patient survival time (time to death) of colon cancer patients by treatment. The curves for observation and Levamisole group show little separation, while Lavamisole with 5-fluorouracil group has a distinct curve. Under the proportional hazards assumption, the Cox regression model can be used to measure treatment effect. If treatment is coded by $(Z_1, Z_2) = (1, 0)$: Levamisole, $(Z_1, Z_2) = (0, 1)$: 5-FU + Levamisole, $(Z_1, Z_2) = (0, 0)$: Observation, then in the model

$$\lambda(t|Z_1, Z_2) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2),$$

$\lambda_0(t)$ represents the hazard function for survival time while not treated (observation), and β_1 is the log of the hazard ratio of Levamisole treated group; i.e., if $\lambda_1(t) \equiv \lambda(t|(Z_1, Z_2) = (1, 0))$ then, for all t,

$$\lambda_1(t)/\lambda_0(t) = e^{\beta_1}.$$

Similarly, β_2 is the log of the hazard ratio of 5-FU + Levamisole treated group; i.e., if $\lambda_2(t) \equiv \lambda(t|(Z_1, Z_2) = (0,1))$ then, for all t,

$$\lambda_2(t)/\lambda_0(t) = e^{\beta_2}.$$

The maximum partial likelihood estimate for β_1 is $\hat{\beta}_1 = -0.02668$ (se: 0.11030, p-value = 0.80888) and that for β_2 is $\hat{\beta}_2 = -0.97169$ (se: 0.11875, p-value = 0.00175). Under the proportional hazards assumption, a 95% confidence interval for hazard ratio of Levamisole group is (0.7844, 1.2087) and that of 5-FU + Levamisole group is (0.5464, 0.8703), indicating that the death rate on Levamisole group to be 0.9737 that on observation group, but not statistically significant. However, we estimate the death rate on 5-FU + Levamisole group to be 0.6896 that on observation group with a statistical significance. The proportional hazards assumption looks reasonable in this case because the Kaplan-Meier curves of 5-FU + Levamisole group and Observation group do not cross. Even though the curves of Levamisole group and Observation group cross each other, it is because there is no statistical evidence that these two curves are distinguishable. Therefore, we conclude that the proportional hazards assumption holds for this model, and indeed it is well fitted.

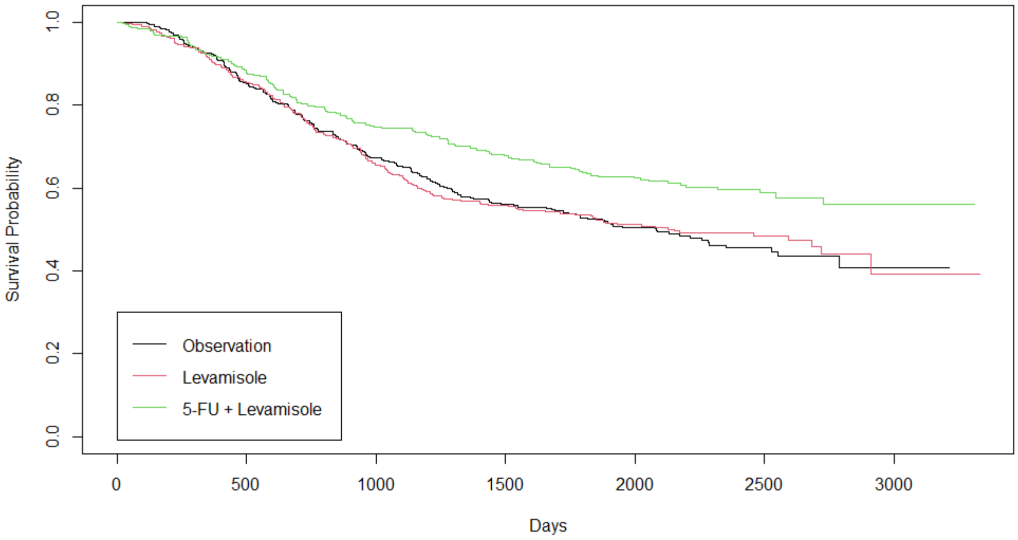

Figure 2: Kaplan-Meier curves of patient survival by treatment groups

We also conducted log rank tests for comparing treatment effects. For comparing Levamisole versus observation, patients in Levamisole and observation groups were used to test the difference between Kaplan-Meier curves (i.e., patients in 5-FU + Levamisole group are not used) using log rank test. Comparing 5-FU + Levamisole versus observation and 5-FU + Levamisole versus Levamisole only were also conducted similarly. Table 2 presents test results, indicating that there is no difference between KM curves of Levamisole and observation group, but there is difference between that of 5-FU + Levamisole group and other groups. Thus, we can conclude that there is evidence of 5-FU + Levamisole has positive effect on decreasing death risk.

Table 2: Log rank tests for comparing treatment effects on patient survival

Comparsion	Chisq test statistic	DF	P-value
Lev vs. Obs	0.057	1	0.8
5-FU + Lev vs. Obs	9.97	1	0.002
5-FU + Lev vs. Lev	8.21	1	0.004

3 Natural history model on patients survival

3.1 Univariate analysis

Table 3: Prognostic Factors: Summary of Univariate Statistics (929 Patients in the colon data set)

	min	1st Q	med	mean	3rd Q	max	Missing	Rao χ^2	DF	pvalue
Pre-operative CEA level	0	2	3	10.75	8	221	442	4.88	1	0.03
Nodal involvemnet	0	1	2	3.66	5	33	18	110.1	1	<2e-16
Age	18	53	61	59.75	69	85	0	0.24	1	0.6
Sex		male: 484			female: 445		0	0.02	1	0.9
Location of primary neoplasm				1: 198 2: 110 3: 52 4: 63 5: 38 6: 49 7: 314 8: 83 9: 0 10: 22			0	13.6	9	0.09
Histologic type				1: 839 2: 78 3: 7 4: 5			0	2.81	3	0.4
Extent of local spread				1: 21 2: 106 3: 759 4: 43			0	29.21	3	6e-06
Differentiation				1: 93 2: 662 3:150			23	17.31	3	6e-04
		Absent			Present		Missing	Rao χ^2	DF	pvalue
Obstruction		749			180		0	5.35	1	0.02
Perforation		902			27		0	0.35	1	0.6
Adherence		794			135		0	6.52	1	0.01
Regional Implants		872			57		0	9.07	1	0.003

The data can be used to build a statistical model for the influence of covariates on survival time. Table 3 provides the distribution of 12 clinical variables as well as their score statistics based on univariate Cox proportional hazard model (i.e., using only one covariate). Nodal involvement was the strongest univariate predictor of survival.

3.2 Variable Selection

Parsimonious but accurate models based on inexpensive, non-invasive and readily available measurements are useful in clinical science, so the variables Pre-operative CEA level, Location of primary neoplasm, Histologic type, and Perforation were eliminated temporarily from the variable selection process. The untransformed versions of the remaining 8 variables were inserted into Cox proportional hazard model, and a step-down procedure was employed to eliminate variables, using the p-value of likelihood ratio test as a criterion for deletion of the least predictive variable (i.e., exclude if p-value $>$ 0.05). Table 4 displays the first step of the procedure, which led to the elimination of the variable sex, and the fifth step, at which each of the remaining variables has a likelihood ratio test p-value smaller than 0.05. Note that even though variable age is not statistically significant in terms of likelihood ratio test p-value $>$ 0.05, since it is obvious that age affects the survival hazard, we retained age in the model. The likelihood ratio test for the four eliminated variables has the value

$$-2(-2791.719 + 2786.540) = 10.358,$$

and has an approximate chi-square distribution with 6 degrees of freedom. There is little evidence to retain the variables Sex, Differentiation, Obstruction, or Adherence.

Table 4: Results of variable selection procedure (929 Patients in the colon data set)

(a) First Step, log likelihood: -2786.540

	Level	Coef	Se	Z stat	Wald p value	Deletion LRT p value
Levamisole		-0.087	0.113	-0.777	0.437	
5-FU + Levamisole		-0.416	0.121	-3.426	0.001	
Sex		**-0.013**	**0.096**	**-0.139**	**0.889**	**0.889**
Differentiation	2	-0.106	0.168	-0.630	0.528	0.218
Differentiation	3	0.164	0.198	0.830	0.406	
Extent of local spread	2	0.670	0.604	1.110	0.267	1.24e-04
Extent of local spread	3	1.188	0.583	2.038	0.042	
Extent of local spread	4	1.604	0.613	2.617	0.009	
Obstruction		0.217	0.118	1.840	0.066	0.071
Adherence		0.189	0.130	1.456	0.145	0.152
Regional implants		0.361	0.181	1.992	0.046	0.057
Nodal involvement		0.087	0.009	9.406	<2e-16	1.35e-15
Age		0.006	0.004	1.555	0.120	

(b) Last Step, log likelihood: -2791.719

	Level	Coef	Se	Z stat	Wald p value	Deletion LRT p value
Levamisole		-0.086	0.112	-0.771	0.441	
5-FU + Levamisole		-0.419	0.121	-3.465	0.001	
Extent of local spread	2	0.718	0.602	1.192	0.233	1.87e-05
Extent of local spread	3	1.258	0.581	2.164	0.030	
Extent of local spread	4	1.745	0.609	2.867	0.004	
Regional implants		0.403	0.181	2.229	0.026	0.034
Nodal involvement		0.089	0.009	9.753	<2e-16	2.24e-16
Age		0.006	0.004	1.549	0.121	

To evaluate the need for transformations of the two continuous variables (Nodal involvement, Age) in the five variable (Treatment, Extent of local spread, Regional implants, Nodal involvement, Age) model in Table 4(b), the variables log(Nodal involvement), log(age) were added. The resulting seven variable model in Table 5(a) provides a significantly better fit than the model in Table 4(b). That is, the likelihood ratio statistic having 2 DF is -2(-2791.719 + 2779.730) = 23.978.

The square transformation of continuous variables were also considered, and the martingale residual plots against the transformation of variables (untransformed, logarithm, and square) are plotted in Figure 3 (Age) and 4 (Nodal involvement). According to the martingale residual plots, log(Age) and log(Nodal involvement) transformation give the best residual distributions. However, since there is no huge difference between the residuals in Age and log(Age), we decided to retain Age because using untransformed variable Age allows much easier interpretation of the coefficient estimate. Table 5(b) presents the log likelihood and regression coefficients for the five variable model containing Treatment, Extent of local spread, Regional implants, log(Nodal involvement), and Age.

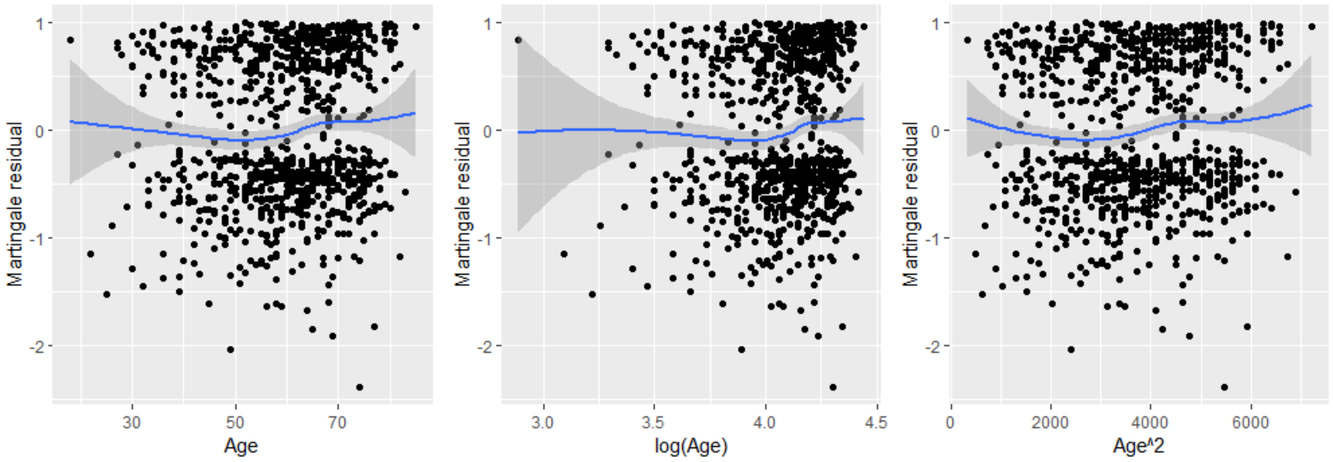

Figure 3: Martingale residual plots against transformations of Age variable

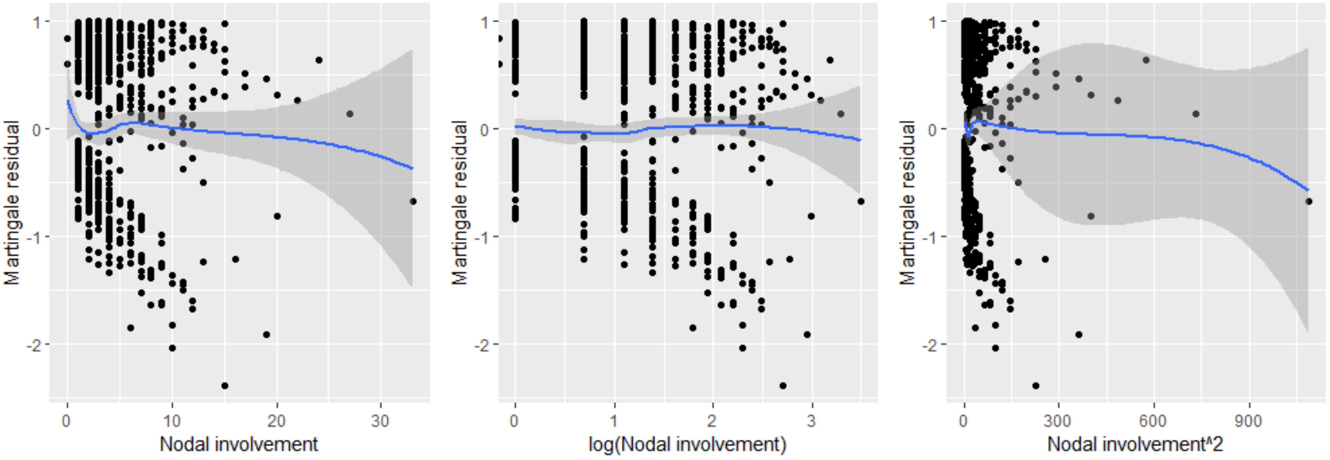

Figure 4: Martingale residual plots against transformations of Nodal involvement variable

The two-way interactions effects of variables were also considered, and interaction effect between Extent of local spread and Age significantly improved model fit (See Table 5(c)) while other two-way interaction effects could not. Therefore, we decided the natural history model including Extent of local spread, Age, and their interaction, Treatment, Regional implants, and logarithm of Nodal involvement. The log likelihood, coefficients, and standard errors for the final model with transformed variables are in Table 5(c).

The Breslow estimate of Λ_0 and estimated survival function

$$\hat{S}(t \mid Z) = \{e^{-\hat{\Lambda}_0(t)}\}^{\exp(\hat{R})},$$

where $R \equiv \beta_1 Z_1 + \cdots + \beta_k Z_k$ is a risk score, provide patient specific survival estimates. For an individual with risk score $\hat{R} = 9.241$, the median risk score in the patients, the corresponding one- and five-year survival estimates are $\hat{S}(365) = 0.928$ and $\hat{S}(1825) = 0.587$.

Consider a low-risk patient with Treatment: 5-FU + Levamisole, Extent of local spread: 1 (submucosa/not muscle), no Regional implants, Nodal involvement: 1, Age: 50 years. Her risk score is 5.437, and her estimated one- and five-year survival estimates are $\hat{S}(365) = 0.998$ and $\hat{S}(1825) = 0.988$.

In a high-risk patient with Treatment: Placebo, Extent of local spread: 4 (contiguous structures), Regional implants exists, Nodal involvement: 33, Age: 50 years, her risk score is 11.813, and her estimated one- and five-year survival estimates are $\hat{S}(365) = 0.379$ and $\hat{S}(1825) = 0.0009$.

Table 5: Regression models with variable transformation and interaction effects

(a) Log of continuous variables, log likelihood: -2779.730

	Level	Coef	Se	Z stat	Wald p value
Levamisole		-0.066	0.112	-0.586	0.558
5-FU + Levamisole		-0.415	0.121	-3.431	0.001
Extent of local spread	2	0.651	0.602	1.080	0.280
Extent of local spread	3	1.149	0.582	1.976	0.048
Extent of local spread	4	1.686	0.609	2.767	0.006
Regional implants		0.345	0.181	1.901	0.057
Nodal involvement		0.002	0.024	0.100	0.920
log(Nodal involvement)		0.544	0.124	4.398	0.000
Age		0.048	0.025	1.966	0.049
log(Age)		-2.192	1.289	-1.701	0.089

(b) Transformation of continuous variables, log likelihood: -2781.082

	Level	Coef	Se	Z stat	Wald p value
Levamisole		-0.077	0.112	-0.692	0.489
5-FU + Levamisole		-0.418	0.121	-3.458	0.001
Extent of local spread	2	0.678	0.602	1.125	0.260
Extent of local spread	3	1.169	0.582	2.009	0.045
Extent of local spread	4	1.692	0.609	2.777	0.005
Regional implants		0.347	0.181	1.917	0.055
log(Nodal involvement)		0.560	0.059	9.560	<2e-16
Age		0.007	0.004	1.801	0.072

(c) Final model, log likelihood: -2777.348

	Level	Coef	Se	Z stat	Wald p value
Levamisole		-0.074	0.112	-0.661	0.509
5-FU + Levamisole		-0.398	0.123	-3.253	0.001
Extent of local spread	2	10.320	5.319	1.940	0.052
Extent of local spread	3	9.392	5.265	1.784	0.074
Extent of local spread	4	10.820	5.383	2.009	0.045
Regional implants		0.350	0.181	1.932	0.053
log(Nodal involvement)		0.567	0.059	9.577	<2e-16
Age		0.138	0.075	1.838	0.066
Extent of local spread : Age	2	-0.152	0.076	-1.992	0.046
Extent of local spread : Age	3	-0.128	0.075	-1.709	0.087
Extent of local spread : Age	4	-0.143	0.077	-1.852	0.064

Finally, adjusting for the variables in Table 5(c), the maximum partial likelihood estimate for $\beta_{\text{Levamisole}}$ is $\hat{\beta}_{\text{Levamisole}} = -0.074$ (se: 0.112, p-value = 0.509) and that for $\beta_{\text{5-FU + Lev}}$ is $\hat{\beta}_{\text{5-FU + Lev}} = -0.398$ (se: 0.123, p-value = 0.001), which were $\hat{\beta}_{\text{Levamisole}} = -0.026$ (se: 0.110, p-value = 0.808) and $\hat{\beta}_{\text{5-FU + Lev}} = -0.971$ (se: 0.118, p-value = 0.001) if unadjusted.

Table 6 represents log rank tests for comparing treatment effects under adjusted covariates effect. Similar to the unadjusted case, there is no difference between KM curves of Levamisole and observation group, but there is difference between that of 5-FU + Levamisole group and other groups. Thus, we can conclude that there is evidence of 5-FU + Levamisole has positive effect on decreasing death risk.

Table 6: Log rank tests for comparing treatment effects on patient survival (under adjusted covariates effect)

Comparsion	Chisq test statistic	DF	P-value
Lev vs. Obs	0.5655	1	0.452
5-FU + Lev vs. Obs	10.093	1	0.001
5-FU + Lev vs. Lev	5.5131	1	0.018

3.3 Model checking

Model checking for the final model was done by various methods. First, martingale residual plots against each variable were investigated, as shown in Figure 5 $\sim$ 9. Residuals are centered around zero for every level of categorical variables and every value of continuous variables, and no extreme residual (absolute residual value $>$ 2.5) was identified for every variable. The variance of residuals over levels of categorical variables were similar except the variable Extent of local spread, but it may be due to the imbalance of data over the levels (most of the patients had Extent of local spread level: Serosa). Overall, model fitting was well done according to the martingale residual plots.

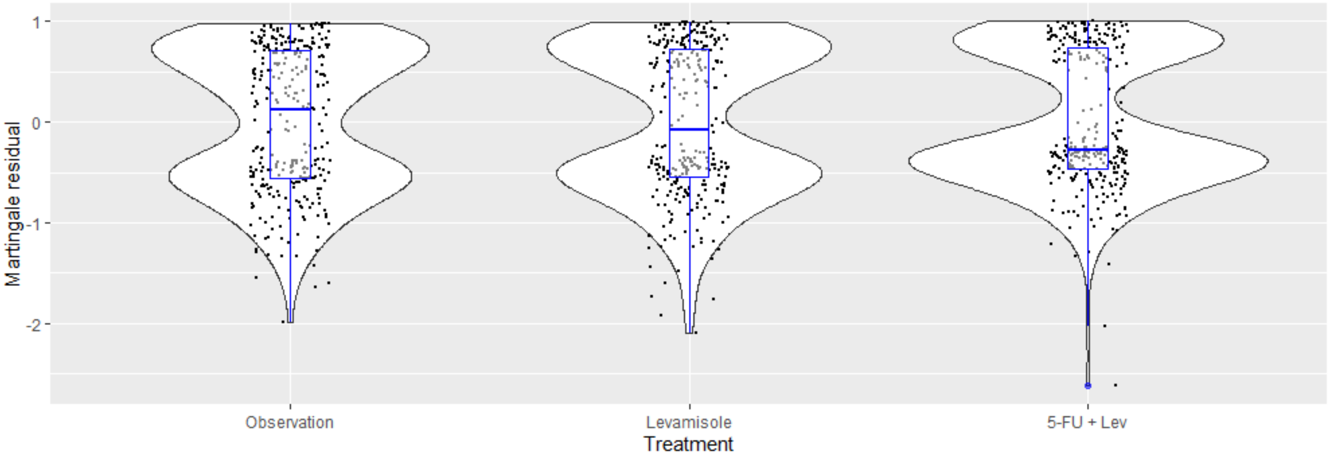

Figure 5: Martingale residual plot of Treatment variable

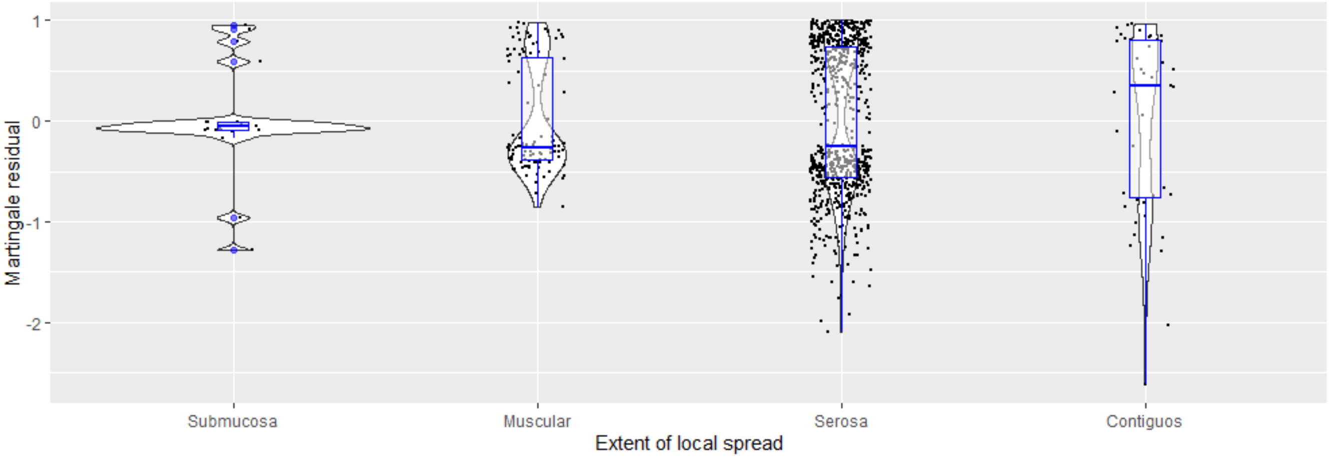

Figure 6: Martingale residual plot of Extent of local spread variable

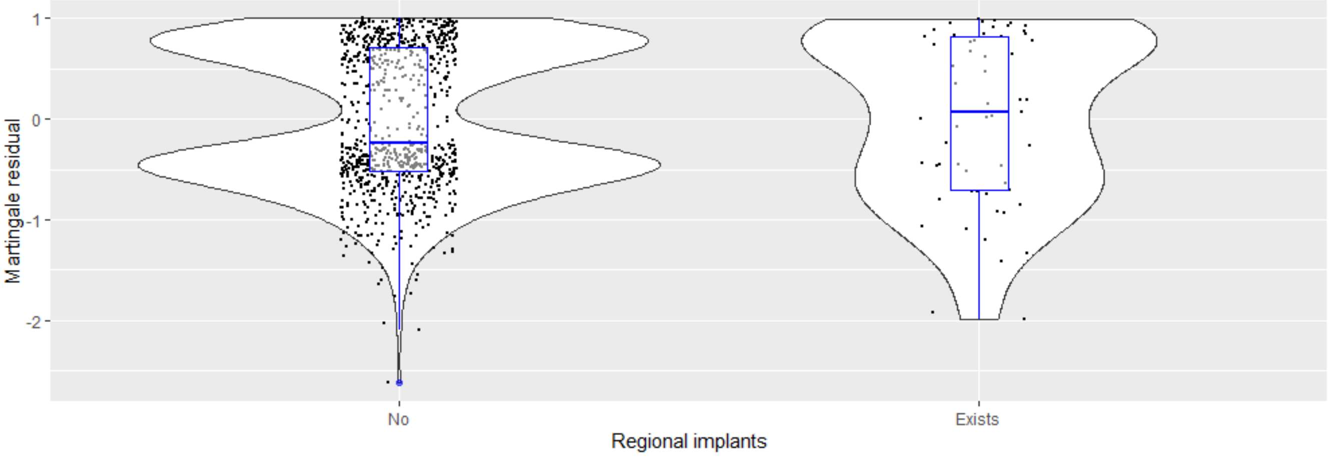

Figure 7: Martingale residual plot of Regional implants variable

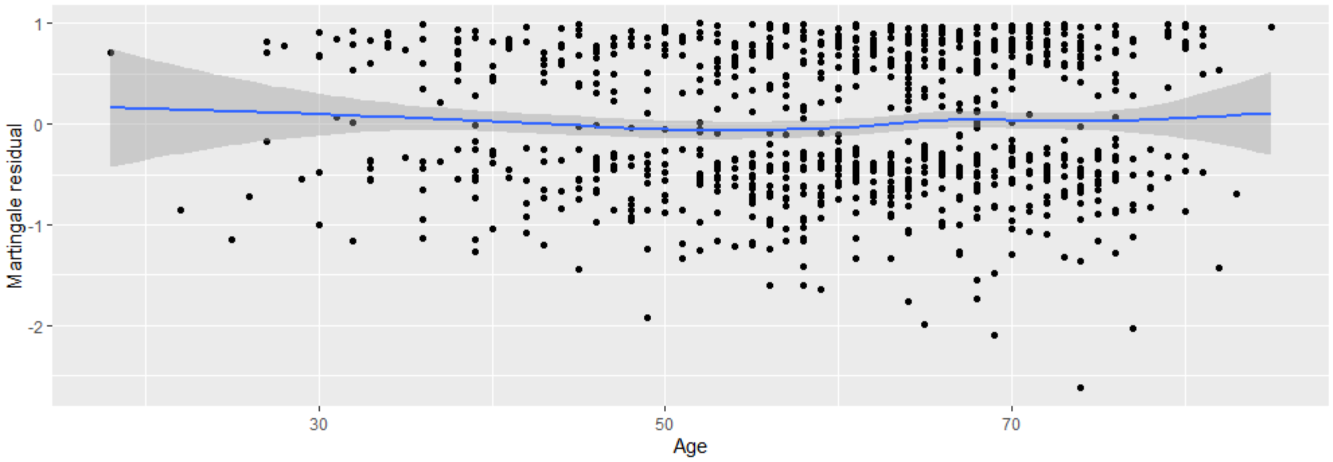

Figure 8: Martingale residual plot of Age variable

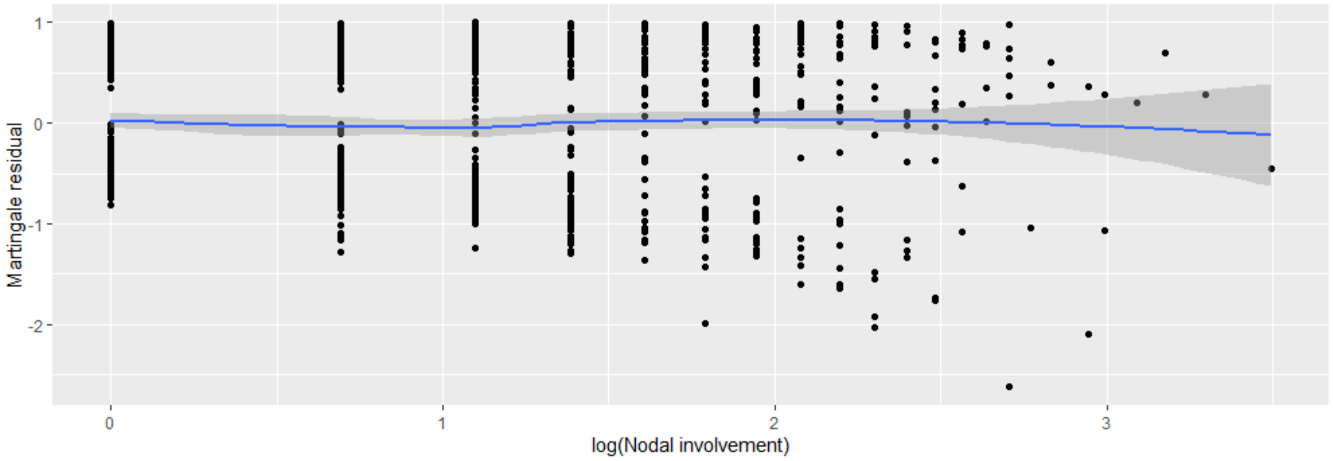

Figure 9: Martingale residual plot of Nodal involvement variable

Next, to assess the functional forms of covariates, cumulative martingale residual plots for Age and logarithm of Nodal involvement were investigated. Kolmogorov-type supremum test were computed on 1000 simulated patterns, and empirical p-values for each covariate were given by 0.410 (Age) and 0.214 (logarithm of Nodal involvement), meaning that each transformation was reasonable (See Figure 10).

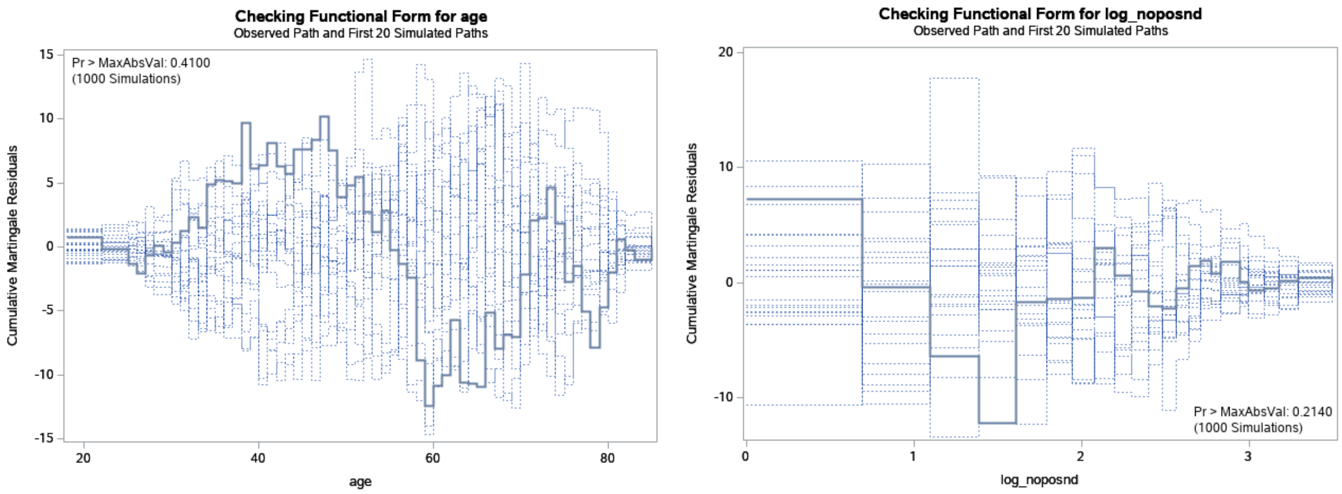

Figure 10: Cumulative martingale residuals plots for Age (Left) and logarithm of Nodal involvement (Right)

Third, proportional hazards assumption for each variable was checked by investigating standardized score process plots, presented at Figure 11. Empirical p-values for each covariate were computed on 1000 simulated patterns, and they were all higher than 0.05, meaning that proportional hazards assumption for each variable is valid.

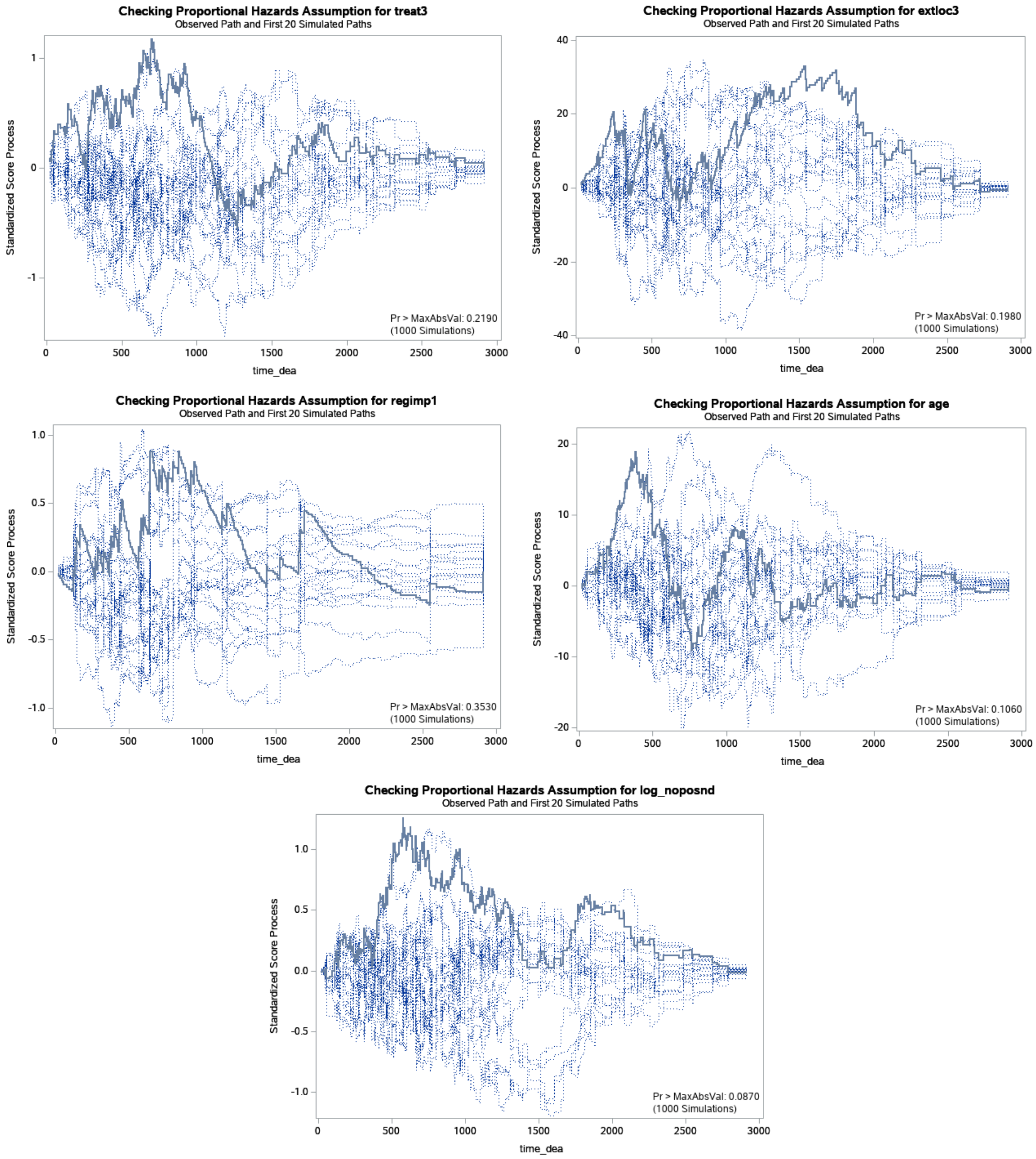

Figure 11: Standardized Score Process plots for Treatment: 5-FU + Lev (Top Left), Extent of local spread: Serosa (Top Right), Regional implants (Middle Left), Age (Middle Right), and logarithm of Nodal involvement (Bottom)

Finally, deviance residual plot (Figure 12, Left) and Schoenfeld residual plot (Figure 12, Right) indicate there is no extreme outlier. In conclusion, the final Cox proportional hazards model containing Extent of local spread, Age, and their interaction, Treatment, Regional implants, and logarithm of Nodal involvement as covariates was well built in terms of model diagnostics.

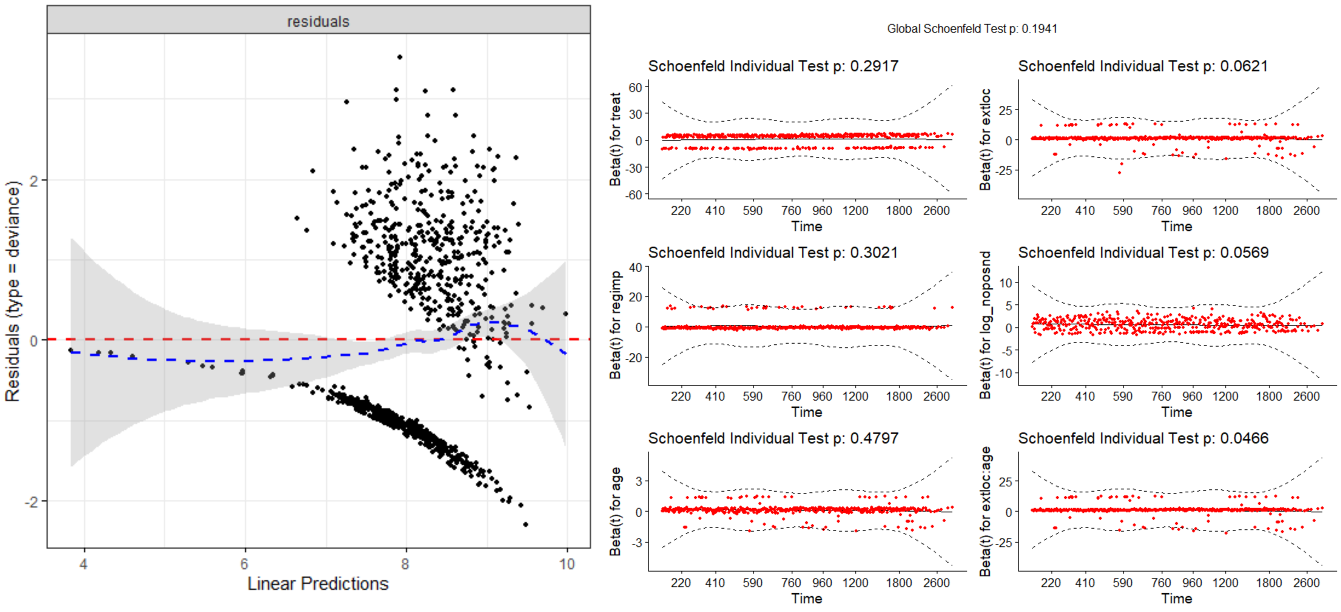

Figure 12: Cumulative martingale residuals plots for Age (Left) and logarithm of Nodal involvement (Right)

3.4 Model validation and Prediction

The natural history model built on the 924 patients from group 1 was assessed using the validation dataset of 243 patients from group 2. Nodal involvement variable was missing for 13 patients, and Regional implants and Extent of local spread variables were missing for all 243 patients in the validation dataset. Nodal involvement variable was imputed by the median of this variable. Regional implants variable was imputed through a logistic regression model on other covariates trained on the model building set (group 1 data) and predicted on the model validation set (group 2 data). Similarly, Extent of local spread variable was imputed through a proportional odds model regression on other covariates. The risk scores then were computed for each of 243 validation patients using $\hat{\beta}$ from Table 5(c). Using these scores, patients were divided into low, medium, and high risk subgroups, using empirical $\frac{1}{3}$- and $\frac{2}{3}$-quantiles as cutoffs . Within each subgroup, the average of the predicted survival curves was compared to the actual survival experience represented by a Kaplan-Meier curve (See Figure 13). Prediction was successful on high risk subgroup, while it were not on the other subgroups. This might be due to the imputation of Regional implants and Extent of local spread variables, which were missing for all 243 patients in the validation dataset.

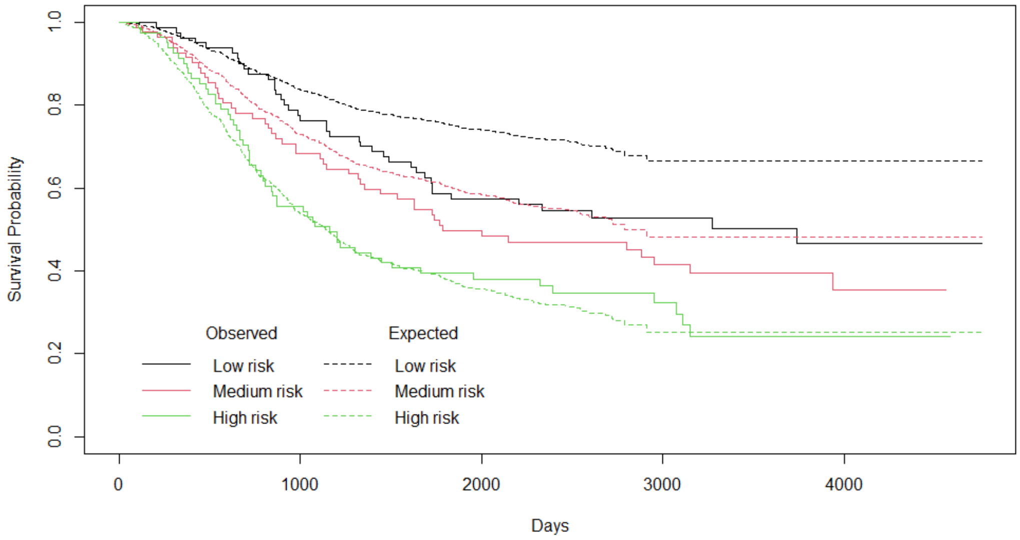

Figure 13: Predicted and observed survival curves by risk group in validation data set

Table 7 compares parameter estimates for the model containing Extent of local spread, Age, and their interaction, Treatment, Regional implants, and logarithm of Nodal involvement when the 929 model building patients, 243 model validating patients, and the total 1172 patients are used in estimation. As seen in the table, the final natural history model was well fitted in the model building dataset, but it was not in the model validating dataset (p-values for Extent of local spread variable are extremely high). Therefore, predicting survival function of validation dataset using the natural history model built on model building dataset cannot be done with high accuracy.

Table 7: Regression Coefficients for Cox Regression Survival Models

		Model Building (n=929)			Model Validating (n=243)			Total Patients (n=1172)		
	Level	Coef	Se	p value	Coef	Se	p value	Coef	Se	p value
Levamisole		-0.074	0.112	0.509	-0.252	0.206	0.222	-0.087	0.097	0.368
5-FU + Levamisole		-0.398	0.123	0.001	-0.387	0.207	0.061	-0.344	0.103	0.001
Extent of local spread	2	10.320	5.319	0.052	-50.370	5.99e+03	0.993	4.832	2.908	0.097
Extent of local spread	3	9.392	5.265	0.074	-48.270	5.99e+03	0.994	4.755	2.826	0.092
Extent of local spread	4	10.820	5.383	0.045	-49.550	5.99e+03	0.993	5.033	2.967	0.090
Regional implants		0.350	0.181	0.053	-0.913	0.393	0.020	0.069	0.158	0.664
log(Nodal involvement)		0.567	0.059	<2e-16	0.487	0.116	1e-05	0.521	0.052	<2e-16
Age		0.138	0.075	0.066	-1.196	1.50e+02	0.994	0.071	0.044	0.110
Extent of local spread : Age	2	-0.152	0.076	0.046	1.235	1.50e+02	0.993	-0.070	0.046	0.125
Extent of local spread : Age	3	-0.128	0.075	0.087	1.199	1.50e+02	0.994	-0.064	0.044	0.148
Extent of local spread : Age	4	-0.143	0.077	0.064	1.228	1.50e+02	0.994	-0.060	0.047	0.199

4 Conclusion

We showed that while the use of levamisole alone is not effective on both preventing disease recurrence and increasing patient survival time, the combination of levamisole and 5-fluorouracil has a positive effect on both of them. Natural history model built on the group 1 patient data (SWOG 8591) identified Extent of local spread, Age, and their interaction, Treatment, Regional implants, and logarithm of Nodal involvement are predictive of the patient survival.

Each covariates and its functional form was validated using martingale residuals and cumulative martingale residuals plots as well as Kolomogorov-type supremum test. Proportional hazard assumption for each variable was validated by standardized score process plots. In addition, there was no extreme outlier according to the deviance residual plot and Schoenfeld residual plots.

The model was validated using group 2 patient data (NCCTG), by predicting the survival curves of low-, medium-, and high-risk group. The prediction was well done at high-risk group, but not at other groups. It is possibly due to the imputation of missing variables in validation set (Extent of local spread, Regional implants), which were identified as statistically significant in model building procedure.

Overall, the final model coefficients estimates imply some scientific discoveries (See Table 7, Right). First, negative value of $\hat{\beta}_{\text{5-FU + Lev}}$ implies that this combination is effective at decreasing risk. Next, positive value of $\hat{\beta}_{\text{Extent of local spread}}$ implies that risk increases as the extent of local spread becomes larger. Third, positive value of $\hat{\beta}_{\text{Regional implants}}$ implies that risk increases if many tumor implants are found in the abdominal lining in the local region of the tumor. Fourth, age and the number of resected lymph nodes involved with cancer is positively correlated with the risk. Finally, statistically significant interaction effect between Extent of local spread and Age implies that the effect of Age on risk varies by the extent of local spread of the tumor.

Reference

- Thomas R. Fleming and David P. Harrington (2005), *Counting Processes and Survival Analysis*, Wiley Series in Probability and Statistics, Wiley
- Danyu Lin (2021), *BIOS 780: Survival Analysis* lecture notes, Department of Biostatistics, University of North Carolina at Chapel Hill

Computing resources

Data analysis and model fitting procedure were done by R and SAS. Every figure and program code can be found at `https://github.com/chanhwa-lee/BIOS780-Survival-Analysis-Final-Project`.