Intro and EDA

Xueqing Huang(xh2470)

Introduction

Lung cancer is a disease with a very high prevalence. Prognostic factors provide important information for patients with cancer. A better understanding of patients' prognosis can help in making appropriate therapeutic decisions[1]. Driven by the desire to improve life quality of lung cancer patients, we perform a survival analysis of these patients and analyze factors that affect survival time.

The dataset we use is the lung cancer dataset in 'survival' package in R. The data describes survival of patients with advanced lung cancer from the North Central Cancer Treatment Group, as well as measures of the patients performance assessed either by the physician and by the patients themselves[1]. Our project aims to explore whether factors such as age, sex, and caloric intake, will bring significant differences in the survival rate of patients with advanced lung cancer. The association between both the physician's assessments of performance status as well as the patient's assessment of their own performance status and the survival rate are also evaluated.

Methods we use in this project include exploratory data analysis, non-parametric estimate, hypothesis testing, semi-parametric model and parametric models. Details of those methods are given below.

Methods

Exploratory Data Analysis

The dataset contains a total of 228 patients and 10 variables. A brief description of variables in the dataset is shown below.

- inst: Institution code
- time: Survival time in days
- status: Censoring status(1=censored, 2=dead)
- age: Age in years
- sex: Male=1, Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (from bad=0 to good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient
- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in last six months

Survival endpoint is the death of patients. The type of censoring is right censoring, which means patients left the study before their death. Among 228 patients, 63 of them were right censored and the number of events was 165. We group the patients by their survival status and provide the descriptive statistics of other variables. Wilcoxon rank sum test, Pearson's Chi-squared test, and Fisher's exact test were used to compare values across group.

Table 1: **Patient Characteristics**

Variable	**Overall**, N = 228	**Alive**, N = 63	**Death**, N = 165	**p-value**
Survival Time (days)	305 (211)	363 (221)	283 (203)	0.003
Age	62 (9)	60 (10)	63 (9)	0.053
Sex				<0.001
Male	138 (61%)	26 (41%)	112 (68%)	
Female	90 (39%)	37 (59%)	53 (32%)	
ECOG Score				0.003
Asymptomatic	63 (28%)	26 (41%)	37 (23%)	
Symptomatic but completely ambulatory	113 (50%)	31 (49%)	82 (50%)	
In bed <50% of the day	50 (22%)	6 (9.5%)	44 (27%)	
In bed > 50% of the day but not bedbound	1 (0.4%)	0 (0%)	1 (0.6%)	
Bedbound	0 (0%)	0 (0%)	0 (0%)	
Missing	1	0	1	
Karnofsky Score(by physician)				0.057
50	6 (2.6%)	1 (1.6%)	5 (3.0%)	
60	19 (8.4%)	3 (4.8%)	16 (9.8%)	
70	32 (14%)	3 (4.8%)	29 (18%)	
80	67 (30%)	20 (32%)	47 (29%)	
90	74 (33%)	25 (40%)	49 (30%)	
100	29 (13%)	11 (17%)	18 (11%)	
Missing	1	0	1	
Karnofsky Score(by patients)				0.043
30	2 (0.9%)	1 (1.6%)	1 (0.6%)	
40	2 (0.9%)	1 (1.6%)	1 (0.6%)	
50	4 (1.8%)	0 (0%)	4 (2.5%)	
60	30 (13%)	3 (4.8%)	27 (17%)	
70	41 (18%)	10 (16%)	31 (19%)	
80	51 (23%)	12 (19%)	39 (24%)	
90	60 (27%)	22 (35%)	38 (23%)	
100	35 (16%)	14 (22%)	21 (13%)	
Missing	3	0	3	
Calories Consumed (kcals)	929 (402)	913 (453)	934 (384)	0.4
Missing	47	16	31	
Weight Loss (pounds)	10 (13)	9 (13)	10 (13)	0.3
Missing	14	1	13	

From the table, we can see that average survival time for censored and dead patients is 363 days and 283 days, respectively. From the p values, we can see that for patients who were alive and dead, the survival time, sex proportion, ECOG performance score and Karnofsky performance score rated by patient are significantly different. However, there are no significant differences in age, Karnofsky performance score rated by physician, calories consumed, and weight loss.

From the table we can see that there are some missing values in this dataset. For simplicity, we removed those missing data for the following analysis.

Conclusion

Discussion

Reference

[1] Loprinzi CL. Laurie JA. Wieand HS. Krook JE. Novotny PJ. Kugler JW. Bartel J. Law M. Bateman M. Klatt NE. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*. 12(3):601-7, 1994.