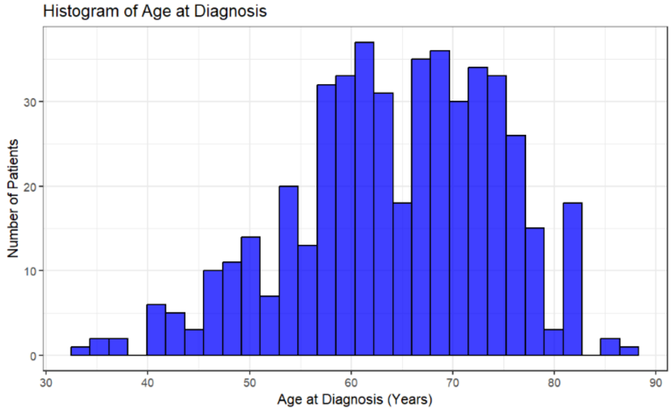
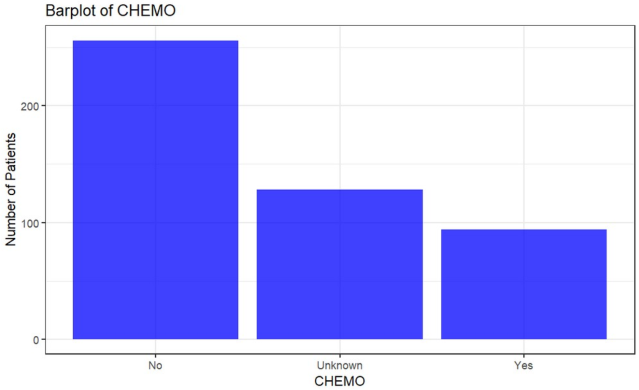
About the Dataset:

The dataset is from the National Cancer Institute's Director's Challenge Lung Study. This data consists of lung cancer data from individuals drawn from 4 different cancer centers. The adjuvant chemo is our covariate of interest.

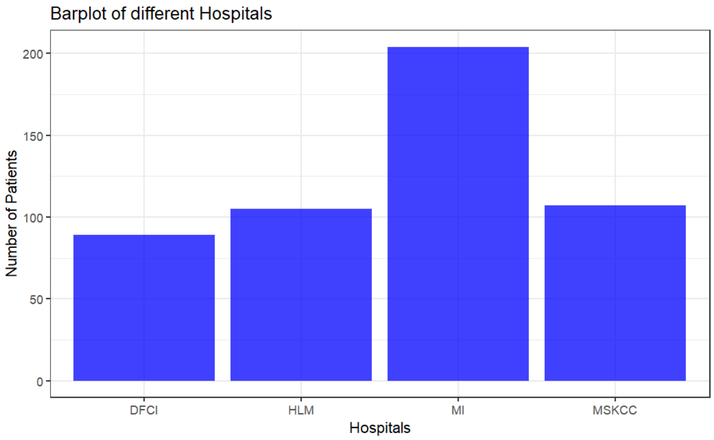
Descriptive Statistics:



Age of Diagnosis seems to be fairly symmetrically distributed and range from around 30 to 90 years.

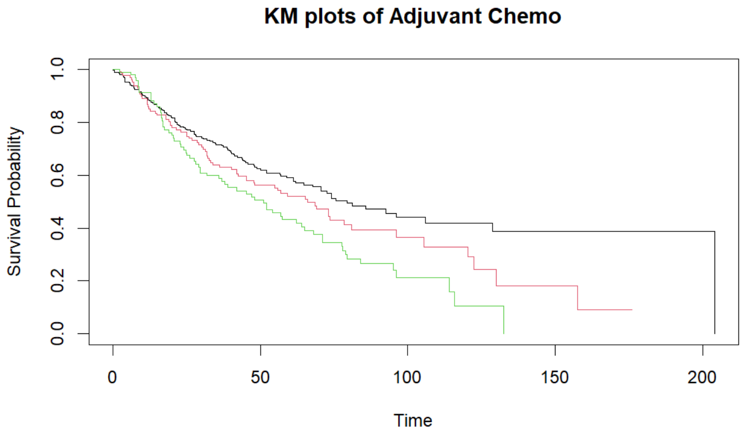


Here we see that there are more patients who did not have to take chemotherapy than those who did.



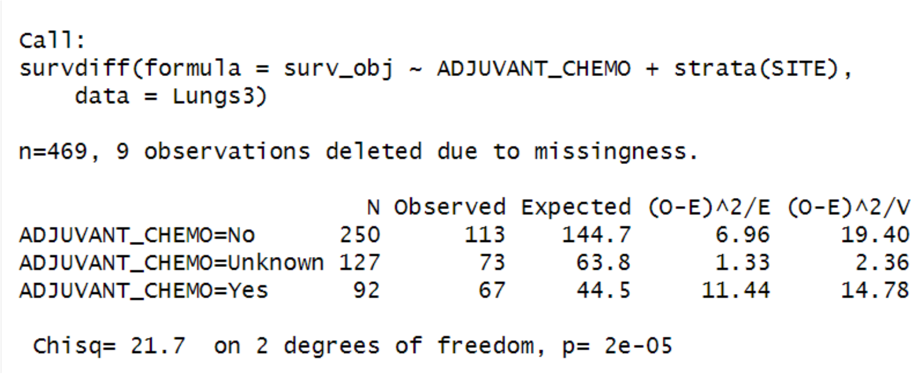
Here we see that there are more patients in MI hospitals, over 200 patients with lung cancer. On the other hand, DFCI hospital has less than 100 patients with lung cancer. While the 2 hospitals have a bit more than 100 each.

Kaplan Meier Plots



Here the black line represents the group who received no chemotherapy. While the red group is unknown, and the green line represents the group who had received chemo. We see that people who did not need chemotherapy survived longer as expected since, they were much healthier than those who needed chemo. Since we can clearly see the separation of the 3 lines, the difference seems significant.

Log-Rank test

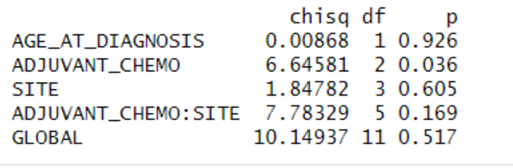


```
Call:
survdiff(formula = surv_obj ~ ADJUVANT_CHEMO + strata(SITE),
    data = Lungs3)

n=469, 9 observations deleted due to missingness.

                         N Observed Expected (O-E)^2/E (O-E)^2/V
ADJUVANT_CHEMO=No       250      113    144.7      6.96     19.40
ADJUVANT_CHEMO=Unknown 127       73     63.8      1.33      2.36
ADJUVANT_CHEMO=Yes       92       67     44.5     11.44     14.78

 Chisq= 21.7  on 2 degrees of freedom, p= 2e-05
```

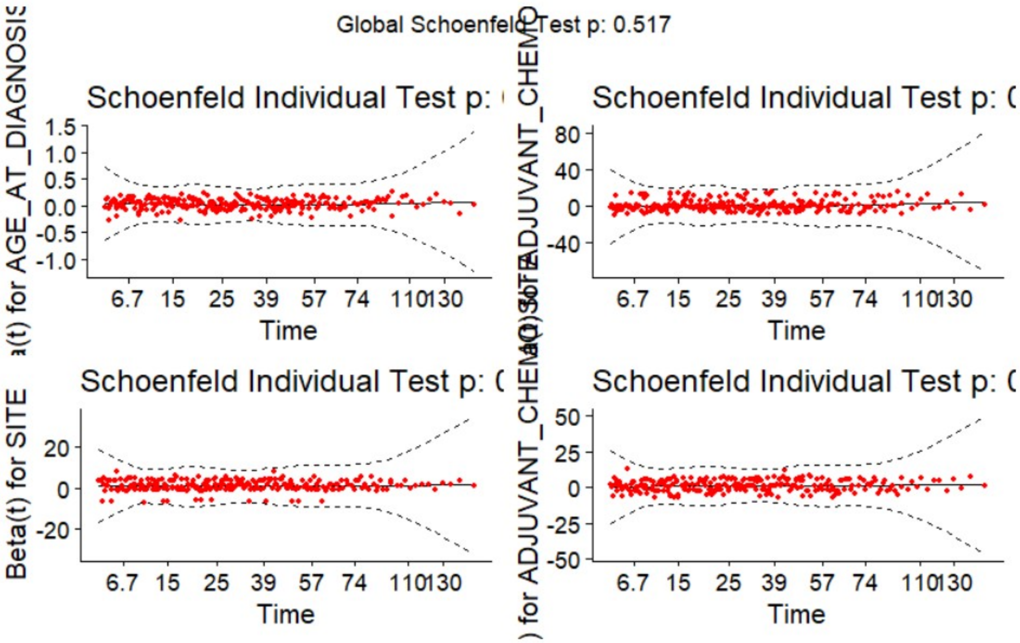
Distribution: Chi-square with 2 degrees of freedom. The p-value: 2e-05, so we have sufficient evidence to reject the null hypothesis. There is a significant effect on chemotherapy in survival stratified by different hospitals.

Cox-PH model:

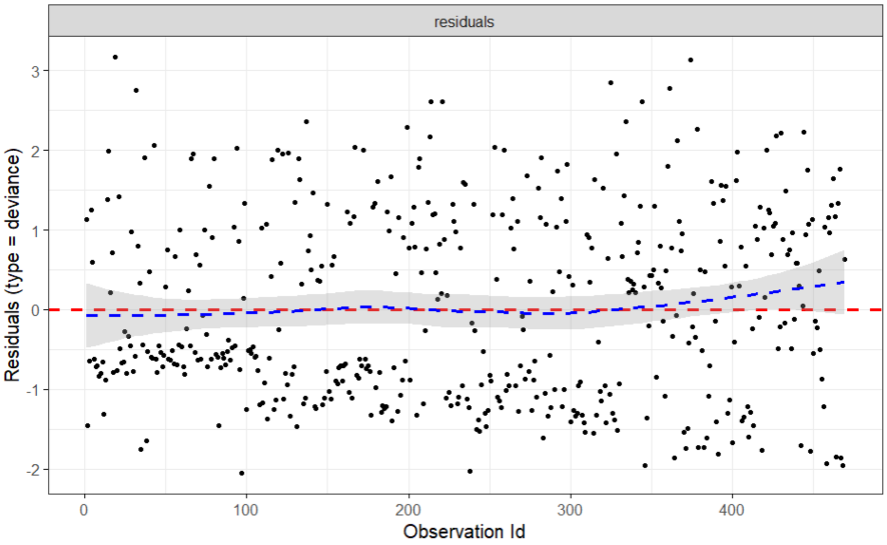
In the survival object, we choose the month to last contact or death as events, while vital status were chosen as censor. Covariate of interest was chemotherapy, and hospital site. The interaction term between these 2 variables were also considered. We choose age as a confounding variable in our model. From the Schoenfeld residuals output below, the test is not statistically significant for each of the covariates, and the global test is also not statistically significant. Therefore, we can assume the proportional hazards.



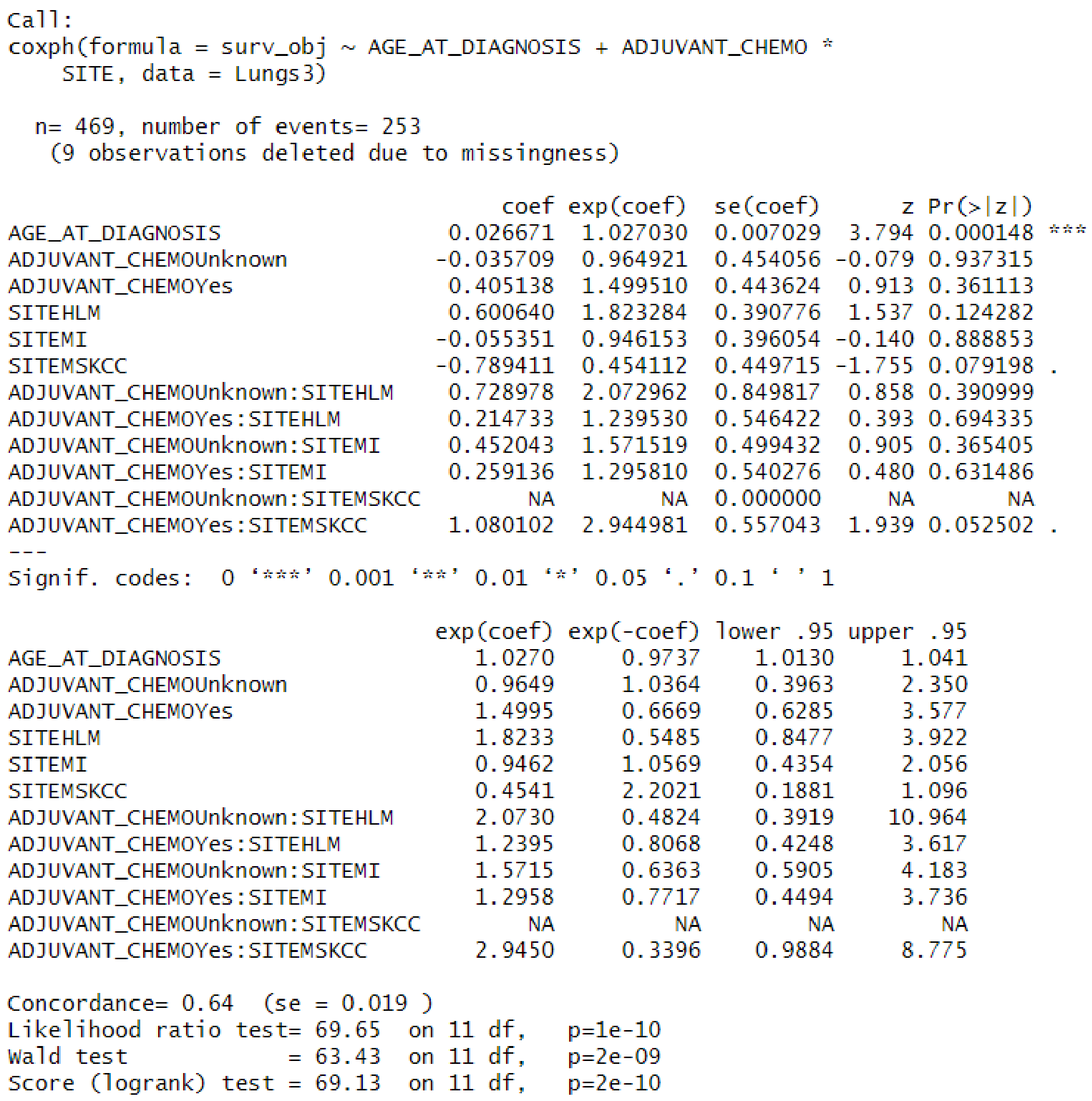
```
                        chisq df     p
AGE_AT_DIAGNOSIS      0.00868  1 0.926
ADJUVANT_CHEMO        6.64581  2 0.036
SITE                  1.84782  3 0.605
ADJUVANT_CHEMO:SITE   7.78329  5 0.169
GLOBAL               10.14937 11 0.517
```



There is no pattern with time. The assumption of proportional hazards appears to be supported for the covariates.



The pattern looks fairly symmetric around 0.



```
Call:
coxph(formula = surv_obj ~ AGE_AT_DIAGNOSIS + ADJUVANT_CHEMO *
    SITE, data = Lungs3)

  n= 469, number of events= 253
   (9 observations deleted due to missingness)

                                       coef exp(coef)  se(coef)      z Pr(>|z|)
AGE_AT_DIAGNOSIS                   0.026671  1.027030  0.007029  3.794 0.000148 ***
ADJUVANT_CHEMOUnknown             -0.035709  0.964921  0.454056 -0.079 0.937315
ADJUVANT_CHEMOYes                  0.405138  1.499510  0.443624  0.913 0.361113
SITEHLM                            0.600640  1.823284  0.390776  1.537 0.124282
SITEMI                            -0.055351  0.946153  0.396054 -0.140 0.888853
SITEMSKCC                         -0.789411  0.454112  0.449715 -1.755 0.079198 .
ADJUVANT_CHEMOUnknown:SITEHLM      0.728978  2.072962  0.849817  0.858 0.390999
ADJUVANT_CHEMOYes:SITEHLM          0.214733  1.239530  0.546422  0.393 0.694335
ADJUVANT_CHEMOUnknown:SITEMI       0.452043  1.571519  0.499432  0.905 0.365405
ADJUVANT_CHEMOYes:SITEMI           0.259136  1.295810  0.540276  0.480 0.631486
ADJUVANT_CHEMOUnknown:SITEMSKCC          NA        NA  0.000000     NA       NA
ADJUVANT_CHEMOYes:SITEMSKCC        1.080102  2.944981  0.557043  1.939 0.052502 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                  exp(coef) exp(-coef) lower .95 upper .95
AGE_AT_DIAGNOSIS                     1.0270     0.9737    1.0130     1.041
ADJUVANT_CHEMOUnknown                0.9649     1.0364    0.3963     2.350
ADJUVANT_CHEMOYes                    1.4995     0.6669    0.6285     3.577
SITEHLM                              1.8233     0.5485    0.8477     3.922
SITEMI                               0.9462     1.0569    0.4354     2.056
SITEMSKCC                            0.4541     2.2021    0.1881     1.096
ADJUVANT_CHEMOUnknown:SITEHLM        2.0730     0.4824    0.3919    10.964
ADJUVANT_CHEMOYes:SITEHLM            1.2395     0.8068    0.4248     3.617
ADJUVANT_CHEMOUnknown:SITEMI         1.5715     0.6363    0.5905     4.183
ADJUVANT_CHEMOYes:SITEMI             1.2958     0.7717    0.4494     3.736
ADJUVANT_CHEMOUnknown:SITEMSKCC          NA         NA        NA        NA
ADJUVANT_CHEMOYes:SITEMSKCC          2.9450     0.3396    0.9884     8.775

Concordance= 0.64  (se = 0.019 )
Likelihood ratio test= 69.65  on 11 df,   p=1e-10
Wald test            = 63.43  on 11 df,   p=2e-09
Score (logrank) test = 69.13  on 11 df,   p=2e-10
```

Conclusion:

Here we see the age of diagnosis is significant at 5% level of significance. With a hazard ratio of 1.027, so older patients were at more risk of dying before they get treated from their lung cancer. All the other variables were not significant.

Software Use:

R 4.4.1. Library ggplot2 and Survival were used for plotting and doing the analysis respectively.