Final Project

Lihui Chen Xuechen Yu Richard Li (Xuanyou)

1. Background and Motivation

Metastatic colorectal cancer (mCRC), which is a significant challenge in oncology, remains something that has a substantial impact on patient morbidity and mortality. Even though there have been a lot of advancements in therapeutic challenges, the prognosis for patients with mCRC is often poor. The selection of first-line therapy is thus a critical determinant of clinical outcomes.

The management of mCRC has evolved with the introduction of targeted therapies, which aim to improve survival rates and quality of life. Among these, the epidermal growth factor receptor (EGFR) inhibitor panitumumab has shown promise when used in combination with the conventional chemotherapy regimen FOLFOX (a combination of folinic acid, fluorouracil, and oxaliplatin).

The trial, NCT00364013, is a randomized, multicenter, phase 3 study. It was designed to evaluate the efficacy of panitumumab in combination with FOLFOX versus FOLFOX alone as first-line therapy for patients with previously untreated mCRC. This study holds a repository of rich clinical data, including patient demographics, treatment details, response criteria, survival metrics, and adverse events. Such comprehensive data presents an invaluable opportunity for the application of advanced statistical and machine learning models to predict patient outcomes.

The motivation behind utilizing these computational models lies in their ability to investigate complex patterns within the data. It potentially leads to the identification of prognostic variables and the development of predictive models for patient survival. By applying these methodologies, the research can shed light on the multifaceted nature of mCRC progression and response to treatment. Moreover, it strives to enhance the decision-making process in clinical settings, enabling personalized medicine approaches and optimizing therapeutic efficacy.

2. Research question

1. How can robust statistical and machine learning models be developed and validated using the trial dataset to accurately predict mortality in patients with metastatic colorectal cancer (mCRC)?
2. What are the key clinical and treatment-related factors that significantly impact the survival of patients with mCRC, and how can their influence be quantified and incorporated into predictive models for patient survival?

3. Data cleaning and exploration

The data are separated into a number of files, and we primarily focused on the adsl and biomark data. We first converted the data type and then imported the file into R studio. The adsl file is full_joined by biomark file, and then we converted the categorical data using one hot encoding and dummy variable. Except the categorial column of sex and race, all the other categorical variables are converted using one hot encoding.

In the data preprocessing phase of our analysis, we focused on a selected set of variables believed to influence patient outcomes in metastatic colorectal cancer. We first isolated these key variables and then quantified the missing values for each. To maintain the integrity of our analysis, we opted for listwise deletion, removing any records with missing data to prevent potential biases that could arise from incomplete information. This resulted in a refined dataset, which was cleansed of missing values and thus poised for the next stage of our investigation. We documented the reduction in the dataset size to ensure transparency in our methodology and the implications it may have on the study's findings.

```
library(bis620.2022)
library(readr)
library(dplyr)
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
library(tidyr)
```

```
data <- adsl %>%
  full_join(biomark, by = 'SUBJID')

final_df_clean = cleaning_helperfunc(data)
#>    Variable MissingValuesCount
#> 1      ATRT                  0
#> 2    PRSURG                  0
#> 3  LIVERMET                  0
#> 4       AGE                  0
#> 5       SEX                  0
#> 6  B_WEIGHT                  0
#> 7  B_HEIGHT                  1
#> 8    B_ECOG                  0
#> 9  B_METANM                  3
#> 10 DIAGTYPE                  0
#> 11      DTH                  0
#> 12    DTHDY                  0
#> [1] "Original number of rows: 935"
#> [1] "Number of rows after removing missing data: 931"
help(cleaning_helperfunc)
# Description
# This function performs several cleaning operations on a data frame. It converts categorical columns t
```

```
head(final_df_clean)
#> # A tibble: 6 x 12
#>    ATRT PRSURG LIVERMET   AGE   SEX B_WEIGHT B_HEIGHT B_ECOG B_METANM DIAGTYPE
#>   <dbl>  <dbl>    <dbl> <dbl> <dbl>    <dbl>    <dbl>  <dbl>    <dbl>    <dbl>
#> 1     0      1        1    64     1     64.2      167      2        3        0
#> 2     1      1        1    65     1     67        165      2        5        0
#> 3     0      1        1    60     1     75        175      2        1        0
#> 4     0      1        1    64     0     52        160      2        2        0
#> 5     0      0        1    74     1     66        171      0        1        1
```

```
#> 6     1      1        1    52     0     53        160      2        3        0
#> # i 2 more variables: DTH <dbl>, DTHDY <dbl>
```

4. Analysis

```
# Relevant Package Installation
suppressPackageStartupMessages(library(randomForest))
suppressPackageStartupMessages(library(rpart.plot))
suppressPackageStartupMessages(library(caTools))
suppressPackageStartupMessages(library(caret))
suppressPackageStartupMessages(library(pROC))
```

4.1 Survival Analysis – Coxph Model

For our statistical analysis, we first applied Cox proportional hazards model which examines a times-to-event outcome, t, as a function of one or more exposure variables, x_i. In our case, the times-to-event outcome is the days to death of each patient enrolled in this case. For right-censoring patients, days for them to exit the cohort study or lose contact is recorded. We employed the Cox proportional hazards regression model to investigate the influence of clinical and demographic factors on the survival of patients with metastatic colorectal cancer. The relative importance of the variables can be manifested by the magnitude of the relative value of the coefficients in the model. In addition, we consider the predictability of the Cox PH model by calculating its concordance score, which is The model incorporated several variables, including primary tumor site (ATRT), prior surgical history (PRSURG), presence of liver metastases (LIVERMET), age (AGE), sex (SEX), baseline weight (B_WEIGHT), baseline height (B_HEIGHT), baseline Eastern Cooperative Oncology Group performance status (B_ECOG), number of metastatic sites at baseline (B_METANM), and diagnostic type (DIAGTYPE).

```
suppressPackageStartupMessages(library(survival))
Model1 = fit_coxph(final_df_clean)
?fit_coxph
# Description
# This function fits a Cox Proportional Hazards model to the provided dataset using specified covariate
```

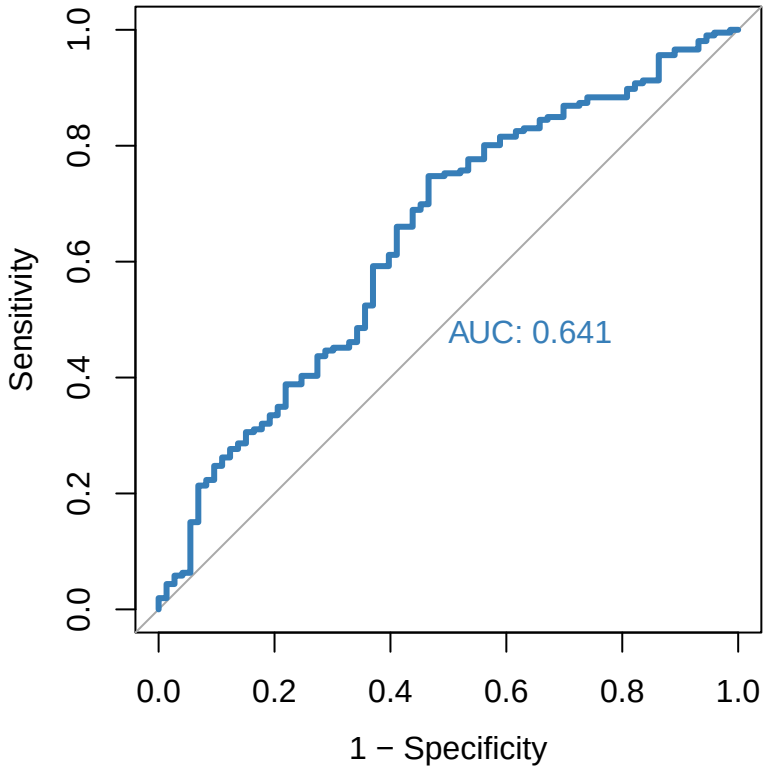
```
Model1
#> Call:
#> coxph(formula = Surv(DTHDY, DTH) ~ ATRT + PRSURG + LIVERMET +
#>     AGE + SEX + B_WEIGHT + B_HEIGHT + B_ECOG + B_METANM + DIAGTYPE,
#>     data = df)
#>
#>               coef exp(coef)  se(coef)      z        p
#> ATRT     -0.008370  0.991665  0.076785 -0.109   0.9132
#> PRSURG   -0.288160  0.749641  0.126651 -2.275   0.0229
#> LIVERMET  0.001244  1.001245  0.122638  0.010   0.9919
#> AGE       0.008882  1.008922  0.003912  2.271   0.0232
#> SEX      -0.099939  0.904893  0.107896 -0.926   0.3543
#> B_WEIGHT -0.007046  0.992978  0.003170 -2.223   0.0262
#> B_HEIGHT  0.009358  1.009402  0.006237  1.500   0.1335
#> B_ECOG    0.217901  1.243464  0.039852  5.468 4.56e-08
#> B_METANM  0.177107  1.193758  0.032586  5.435 5.48e-08
```

```
#> DIAGTYPE -0.223741  0.799522  0.083541 -2.678   0.0074
#>
#> Likelihood ratio test=90.19  on 10 df, p=4.906e-15
#> n= 931, number of events= 687
```

4.2 Machine Learning Model1 – Logistic Regression

In addition to Cox PH model, we also established a logistic regression model, utilizing a machine learning approach to predict the likelihood of death for a certain patient given his or her clinical trail data. There are in total over 931 people, and we did a train test split to leverage 70% of data for training and the other 30% for testing. This approach prevents the model from overfitting and the test accuracy can objectively manifest the predictability of our model. The same set of parameters as in the Cox PH model were used in the logistic regression.

```
suppressPackageStartupMessages(library(caTools))
suppressPackageStartupMessages(library(caret))
suppressPackageStartupMessages(library(pROC))

# Splitting the data into training and test sets
train_test_split <- function(df){

  set.seed(123)
  split <- sample.split(final_df_clean$DTH, SplitRatio = 0.7)
  train_data <- final_df_clean[split == TRUE, ]
  test_data <- final_df_clean[split == FALSE, ]

  return(list(train = train_data, test = test_data))
}

data_all = train_test_split(final_df_clean)
train_data = data_all$train
test_data = data_all$test

Model2 = fit_logistic_regression(train_data, test_data)
#> [1] "Accuracy: 0.745519713261649"
#> Setting levels: control = 0, case = 1
#> Setting direction: controls < cases
```




```
#> [1] "AUC: 0.641308684665514"
#>
#> Call:
#> glm(formula = DTH ~ ATRT + PRSURG + LIVERMET + AGE + SEX + B_WEIGHT +
#>     B_HEIGHT + B_ECOG + B_METANM + DIAGTYPE, family = binomial(),
#>     data = train_data)
#>
#> Coefficients:
#>              Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -3.411249   2.436224  -1.400 0.161448
#> ATRT        -0.221067   0.185295  -1.193 0.232848
#> PRSURG      -0.319157   0.341142  -0.936 0.349502
#> LIVERMET    -0.213059   0.305930  -0.696 0.486160
#> AGE          0.016537   0.008955   1.847 0.064811 .
#> SEX         -0.229159   0.257998  -0.888 0.374422
#> B_WEIGHT    -0.015075   0.007359  -2.048 0.040514 *
#> B_HEIGHT     0.025100   0.015025   1.671 0.094805 .
#> B_ECOG       0.358694   0.101228   3.543 0.000395 ***
#> B_METANM     0.352866   0.091725   3.847 0.000120 ***
#> DIAGTYPE    -0.157325   0.198471  -0.793 0.427961
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 750.34  on 651  degrees of freedom
```

```
#> Residual deviance: 706.58  on 641  degrees of freedom
#> AIC: 728.58
#>
#> Number of Fisher Scoring iterations: 4
?fit_logistic_regression
# Description:
# Fits a logistic regression model to the training data and evaluates its performance on the test data.
```

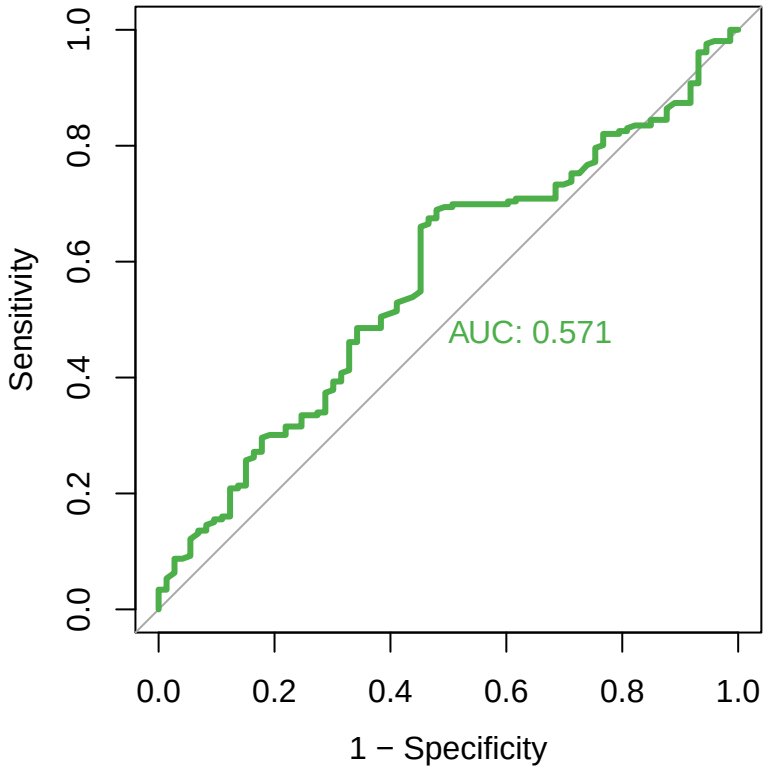
```
Model2
#>
#> Call:  glm(formula = DTH ~ ATRT + PRSURG + LIVERMET + AGE + SEX + B_WEIGHT +
#>     B_HEIGHT + B_ECOG + B_METANM + DIAGTYPE, family = binomial(),
#>     data = train_data)
#>
#> Coefficients:
#> (Intercept)         ATRT       PRSURG     LIVERMET          AGE          SEX
#>    -3.41125     -0.22107     -0.31916     -0.21306      0.01654     -0.22916
#>    B_WEIGHT     B_HEIGHT       B_ECOG     B_METANM     DIAGTYPE
#>    -0.01508      0.02510      0.35869      0.35287     -0.15733
#>
#> Degrees of Freedom: 651 Total (i.e. Null);  641 Residual
#> Null Deviance:       750.3
#> Residual Deviance: 706.6     AIC: 728.6
```

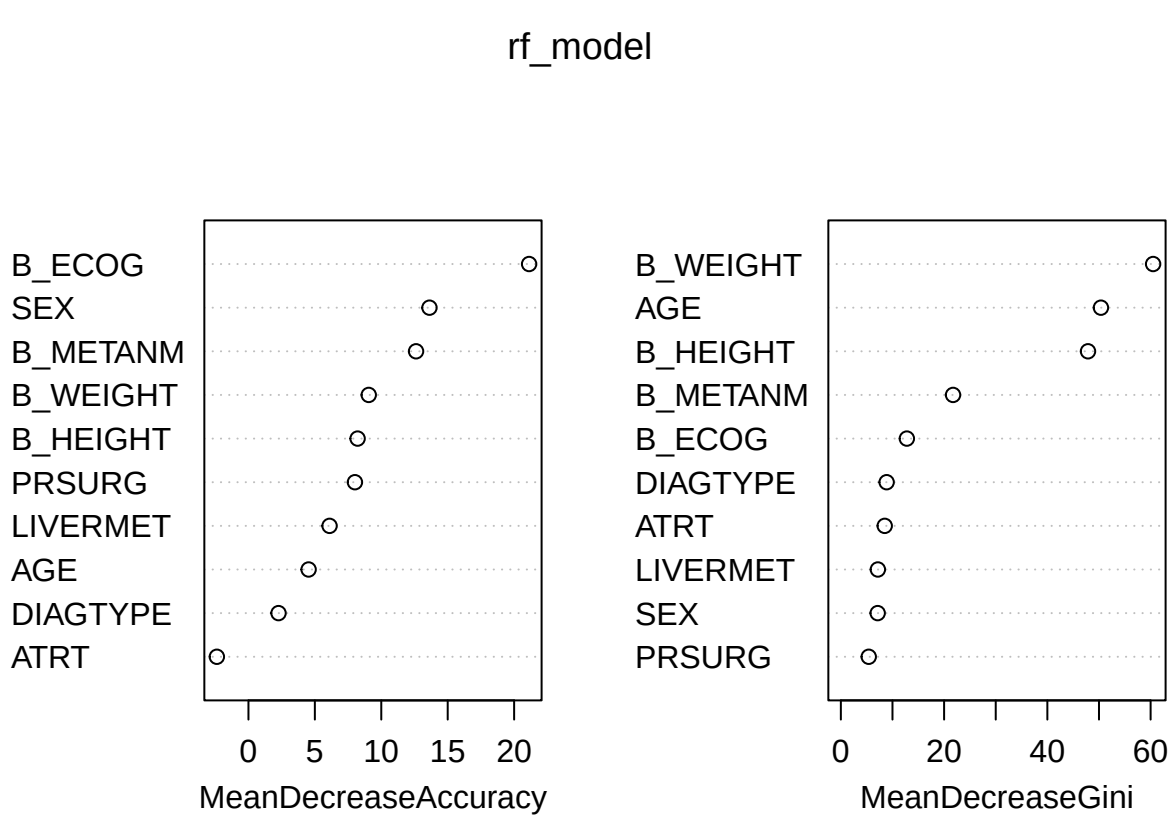
4.3 Machine Learning Model2 – Random Forest

Our group chose random forest as our last machine learning model. This choice has several reasons: 1. The previous two models (Cox PH and Logistic Regression) are both linear models and based on the assumptions for linear models, such as no significant collinearity, etc. We would like to test whether somehow more complicated machine learning model can outperform the linear models. 2. Although random forest is considered a machine learning model, it still preserves some level of interpretability by looking at the features importance. Since this is a clinical trail on patients with colorectal cancer, we do not simply want to accurately predict the probability of death given a patient's condition through a black box approach, we would like to examine the importance of our features as well, which has clincal significance and provide guidance in the treatment and care for the patients.

```
suppressPackageStartupMessages(library(randomForest))
suppressPackageStartupMessages(library(rpart.plot))

# we defined a new function called fit random forest
# The following is our function description:
# Fits a random forest model to the provided training data and evaluates its performance on the test da
?fit_random_forest
```

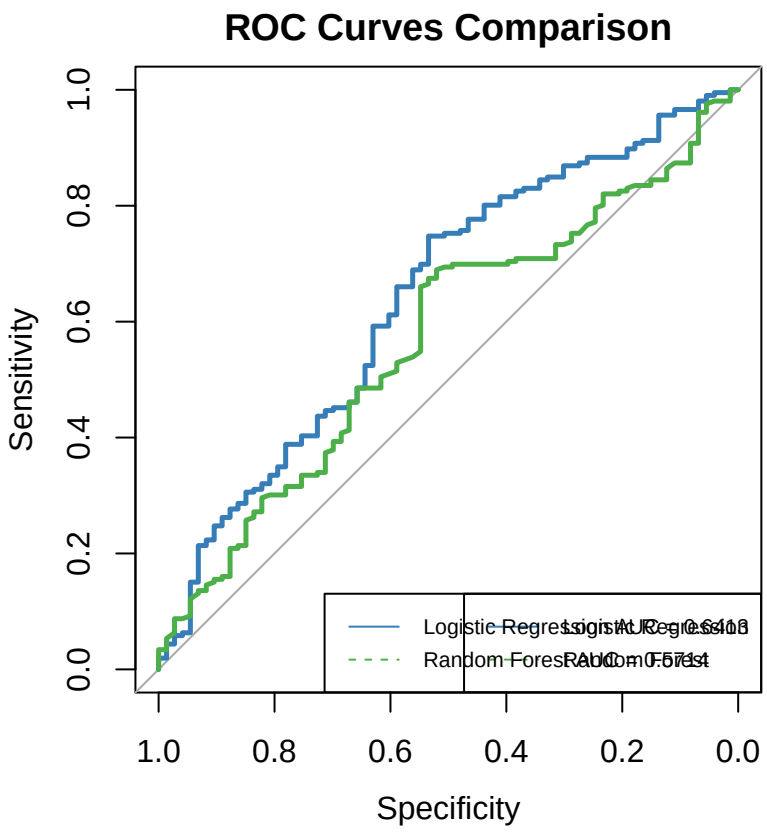
```
Model3 = fit_random_forest(train_data, test_data)
#> [1] "Random Forest model accuracy: 0.666666666666667"
#> Setting levels: control = 0, case = 1
#> Setting direction: controls < cases
```


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 − Specificity
AUC: 0.571



4.4 Visualization

```
generate_roc_for_two <- function(logistic_roc, rf_roc, logistic_auc, rf_auc){

  par(pty = "s")
  plot(logistic_roc, col="#377eb8", lwd = 2.5, main="ROC Curves Comparison")
  # Add ROC curve for random forest
  lines(rf_roc, col="#4daf4a", lwd = 2.5)

  legend("bottomright", legend=c("Logistic Regression", "Random Forest"),
         col=c("#377eb8", "#4daf4a"), lty=c(1, 2), cex=0.7)

  # Optionally, add the AUC scores to the legend if you have them calculated
  # Assume logistic_auc_value and rf_auc_value hold the AUC values for logistic regression and random f
  legend("bottomright", legend=c(paste("Logistic Regression AUC =", round(logistic_auc, 4)),
                                 paste("Random Forest AUC =", round(rf_auc, 4))),
         col=c("#377eb8", "#4daf4a"), lty=c(1, 2), cex=0.7)

  par(pty = "m")

}

generate_roc_for_two(roc_curve, rf_roc_curve, logistic_auc_value, rf_auc_value)
```



5. Interpretation and conclusions

5.1 Results from Cox PH model

```
print(summary(Model1))
#> Call:
#> coxph(formula = Surv(DTHDY, DTH) ~ ATRT + PRSURG + LIVERMET +
#>     AGE + SEX + B_WEIGHT + B_HEIGHT + B_ECOG + B_METANM + DIAGTYPE,
#>     data = df)
#>
#>   n= 931, number of events= 687
#>
#>                coef exp(coef)  se(coef)      z Pr(>|z|)
#> ATRT     -0.008370  0.991665  0.076785 -0.109   0.9132
#> PRSURG   -0.288160  0.749641  0.126651 -2.275   0.0229 *
#> LIVERMET  0.001244  1.001245  0.122638  0.010   0.9919
#> AGE       0.008882  1.008922  0.003912  2.271   0.0232 *
#> SEX      -0.099939  0.904893  0.107896 -0.926   0.3543
#> B_WEIGHT -0.007046  0.992978  0.003170 -2.223   0.0262 *
#> B_HEIGHT  0.009358  1.009402  0.006237  1.500   0.1335
#> B_ECOG    0.217901  1.243464  0.039852  5.468 4.56e-08 ***
#> B_METANM  0.177107  1.193758  0.032586  5.435 5.48e-08 ***
#> DIAGTYPE -0.223741  0.799522  0.083541 -2.678   0.0074 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#>
#>           exp(coef) exp(-coef) lower .95 upper .95
#> ATRT         0.9917     1.0084    0.8531    1.1527
#> PRSURG       0.7496     1.3340    0.5849    0.9609
#> LIVERMET     1.0012     0.9988    0.7873    1.2733
#> AGE          1.0089     0.9912    1.0012    1.0167
#> SEX          0.9049     1.1051    0.7324    1.1180
#> B_WEIGHT     0.9930     1.0071    0.9868    0.9992
#> B_HEIGHT     1.0094     0.9907    0.9971    1.0218
#> B_ECOG       1.2435     0.8042    1.1500    1.3445
#> B_METANM     1.1938     0.8377    1.1199    1.2725
#> DIAGTYPE     0.7995     1.2507    0.6788    0.9418
#>
#> Concordance= 0.615  (se = 0.011 )
#> Likelihood ratio test= 90.19  on 10 df,   p=5e-15
#> Wald test            = 93.84  on 10 df,   p=9e-16
#> Score (logrank) test = 94.93  on 10 df,   p=6e-16
```

As shown in the model summary part, the number of metastatic sites at baseline (B_METANM) had the most significant association with survival rate, with a hazard ratio of 1.194 ($p < 0.001$), which suggests that patients with 1 unit more metastasis were at a 20% higher risk of death. The level of ECOG of a patient, representing the patient's ability of functioning in their daily life (https://ecog-acrin.org/resources/ecog-performance-status/), was also significant in the Cox PH model. A higher score of ECOG means worse physical mobility and the hazard ratio is of 1.243. It means that with one level higher, the patient's at a 24% higher hazard of death. Prior surgical history (PRSURG) and age (AGE) were also significantly associated with survival, with hazard ratios of 0.7496 and 1.009, respectively. The diagnostic type (DIAGTYPE) in addition a significant relationship with a hazard ratio of 0.79952, suggesting a better situation in rectal cancer compared to colon cancer, which is supported by the study of Lee et. al., 2013 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3827090/).

However, the actual treatment the patient receives, whether to be FOLFOX alone or FOLFOX + Panitumumab (ATRT), did not appear to be a significant factor in predicting the patient's death. The presence of liver metastases (LIVERMET), sex (SEX), and baseline height (B_HEIGHT) were also not significantly associated with survival outcomes in this model

The model's concordance statistic was 0.615, indicating a moderate predictive ability. The likelihood ratio test, Wald test, and Score (logrank) test all yielded highly significant p-values ($p < 0.001$), confirming that the model as a whole was statistically significant in distinguishing between different survival times of patients.

5.2 Results from Logistic Regression

```
print(summary(Model2))
#>
#> Call:
#> glm(formula = DTH ~ ATRT + PRSURG + LIVERMET + AGE + SEX + B_WEIGHT +
#>     B_HEIGHT + B_ECOG + B_METANM + DIAGTYPE, family = binomial(),
#>     data = train_data)
#>
#> Coefficients:
#>              Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -3.411249   2.436224  -1.400 0.161448
#> ATRT        -0.221067   0.185295  -1.193 0.232848
```

```
#> PRSURG      -0.319157   0.341142  -0.936 0.349502
#> LIVERMET    -0.213059   0.305930  -0.696 0.486160
#> AGE          0.016537   0.008955   1.847 0.064811 .
#> SEX         -0.229159   0.257998  -0.888 0.374422
#> B_WEIGHT    -0.015075   0.007359  -2.048 0.040514 *
#> B_HEIGHT     0.025100   0.015025   1.671 0.094805 .
#> B_ECOG       0.358694   0.101228   3.543 0.000395 ***
#> B_METANM     0.352866   0.091725   3.847 0.000120 ***
#> DIAGTYPE    -0.157325   0.198471  -0.793 0.427961
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 750.34  on 651  degrees of freedom
#> Residual deviance: 706.58  on 641  degrees of freedom
#> AIC: 728.58
#>
#> Number of Fisher Scoring iterations: 4
```

The final model's accuracy on the test set was reported as approximately 74.55%, which indicates a decent ability to correctly classify the outcomes.

In the analysis of model coefficients, we observed a similar behavior in the Cox PH model but with slight differences in some minor features. In sum, B_ECOG and B_METANM also emerged as significant predictors with $p < 0.001$. Their positive coefficients suggesting that higher values of these variables increase the log-odds of the death of a patient death. However, the other variables, except for weights, were not significant in the logistic regression model. This may because logistic regression only takes the outcome of patient but ignores the time-to-event, day to death in this case.

We also generated the Receiver Operating Characteristic (ROC) curve and computed Area Under the Curve (AUC) to evaluate the discriminative of our model. For logistic regression, an AUC value of 0.641 suggests that the model has a moderate ability to discriminate between patients with outcome and those without.

5.3 Results from Random Forest

```
Model3
#>
#> Call:
#>  randomForest(formula = DTH ~ ATRT + PRSURG + LIVERMET + AGE +      SEX + B_WEIGHT + B_HEIGHT + B_EC
#>                Type of random forest: classification
#>                      Number of trees: 1000
#> No. of variables tried at each split: 3
#>
#>         OOB estimate of  error rate: 27.3%
#> Confusion matrix:
#>    0   1 class.error
#> 0 27 144  0.84210526
#> 1 34 447  0.07068607
```

We also evaluated the performance of a random forest model and obtained a 66.67% of accuracy in the test data. In comparison to linear models used above, random forest model can capture complex interactions

and nonlinear relationships between variables. It also uses bagging strategy i.e. voting from thousands of invididual decision tress to void overfitting.

The variable importance was acquired during the training process of random forest, and it ranks the input variables based on their contribution to the model's predictive power. In our model, the B_ECOG and B_METANM again remained significant among our variables, as indicated by their high Mean Decrease in Accuracy and Gini values. Sex, weight, were also tested important by these two measures, but their actual influence should be further verified.

We also generated ROC curve and computed AUC for random forest, and stacked the plot with logistic regression for easy comparison. A of 0.571 was obtained in random forest, which indicated there is only slight improvement compared to random guess in the random forest model. It also performed worse than logistic regression, which may be attribute to a insufficient tuning of the hyper-parameters.

6. Conclusion

In this panitumumab study, we established three statistical models to investigate their predictability in the patient's death outcome and researched our variables importance through those approaches. Three of our model agreed that the number of metastatis and the level of patient's ability to carry out daily life are sigficant in predicting their death outcome. Yet, other variables such as sex, weight, prior surgery, and cancer type (colon vs rectal), appeared to be significant in some models while lost the significance in others. Their effect should be further and carefully studied.