

Survival Analysis Final Project

Group 10

```
library(readr)
library(tidyverse)
library(janitor)
library(ggcorrplot)
library(dplyr)
library(gtsummary)
```

EDA

Load the dataset and clean column names

```
file_path <- "Breast Cancer METABRIC.csv"

data <- read.csv(file_path, na.strings = "") %>% # Treat blank strings as NA
  clean_names() %>%
  filter(cancer_type != "Breast Sarcoma") %>% # only 3 patients are not Breast Cancer, exclude them
  rename(
    vital_status = patient_s_vital_status,
    three_gene_classifier_subtype = x3_gene_classifier_subtype
  ) %>%
  # Exclude patients with both overall_survival_months and overall_survival_status missing
  filter(!is.na(overall_survival_months) & is.na(overall_survival_status))

# View the cleaned data
head(data)
```

```
##   patient_id age_at_diagnosis type_of_breast_surgery cancer_type
## 1      MB-0000          75.65        Mastectomy Breast Cancer
## 2      MB-0002          43.19        Breast Conserving Breast Cancer
## 3      MB-0005          48.87        Mastectomy Breast Cancer
## 4      MB-0006          47.68        Mastectomy Breast Cancer
## 5      MB-0008          76.97        Mastectomy Breast Cancer
## 6      MB-0010          78.77        Mastectomy Breast Cancer
##
##                               cancer_type_detailed cellularity chemotherapy
## 1             Breast Invasive Ductal Carcinoma      <NA>       No
## 2             Breast Invasive Ductal Carcinoma       High       No
## 3             Breast Invasive Ductal Carcinoma       High      Yes
## 4 Breast Mixed Ductal and Lobular Carcinoma     Moderate      Yes
## 5 Breast Mixed Ductal and Lobular Carcinoma       High      Yes
## 6             Breast Invasive Ductal Carcinoma     Moderate       No
##   pam50_claudin_low_subtype cohort er_status_measured_by_ihc er_status
## 1           claudin-low         1            Positve  Positive
```

```

## 2 LumA 1 Positve Positive
## 3 LumB 1 Positve Positive
## 4 LumB 1 Positve Positive
## 5 LumB 1 Positve Positive
## 6 LumB 1 Positve Positive
## neoplasm_histologic_grade her2_status_measured_by_snp6 her2_status
## 1 3 Neutral Negative
## 2 3 Neutral Negative
## 3 2 Neutral Negative
## 4 2 Neutral Negative
## 5 3 Neutral Negative
## 6 3 Neutral Negative
## tumor_other_histologic_subtype hormone_therapy inferred_menopausal_state
## 1 Ductal/NST Yes Post
## 2 Ductal/NST Yes Pre
## 3 Ductal/NST Yes Pre
## 4 Mixed Yes Pre
## 5 Mixed Yes Post
## 6 Ductal/NST Yes Post
## integrative_cluster primary_tumor_laterality lymph_nodes_examined_positive
## 1 4ER+ Right 10
## 2 4ER+ Right 0
## 3 3 Right 1
## 4 9 Right 3
## 5 9 Right 8
## 6 7 Left 0
## mutation_count nottingham_prognostic_index oncotree_code
## 1 NA 6.044 IDC
## 2 2 4.020 IDC
## 3 2 4.030 IDC
## 4 1 4.050 MDLC
## 5 2 6.080 MDLC
## 6 4 4.062 IDC
## overall_survival_months overall_survival_status pr_status radio_therapy
## 1 140.50000 Living Negative Yes
## 2 84.63333 Living Positive Yes
## 3 163.70000 Deceased Positive No
## 4 164.93333 Living Positive Yes
## 5 41.36667 Deceased Positive Yes
## 6 7.80000 Deceased Positive Yes
## relapse_free_status_months relapse_free_status sex
## 1 138.65 Not Recurred Female
## 2 83.52 Not Recurred Female
## 3 151.28 Recurred Female
## 4 162.76 Not Recurred Female
## 5 18.55 Recurred Female
## 6 2.89 Recurred Female
## three_gene_classifier_subtype tumor_size tumor_stage vital_status
## 1 ER-/HER2- 22 2 Living
## 2 ER+/HER2- High Prolif 10 1 Living
## 3 <NA> 15 2 Died of Disease
## 4 <NA> 25 2 Living
## 5 ER+/HER2- High Prolif 40 2 Died of Disease
## 6 ER+/HER2- High Prolif 31 4 Died of Disease

```

Our primary objective in this project is to investigate factors associated with overall survival (OS) in the METABRIC breast cancer cohort. Specifically, we define overall survival as the time from diagnosis to death from any cause, with overall_survival_months providing the follow-up time and overall_survival_status indicating whether the death event occurred.

For the survival analysis, both a valid survival time and an event indicator are required for each patient. In this dataset, some patients have overall_survival_months and overall_survival_status missing simultaneously, meaning that no follow-up or outcome information is available for these individuals. These cases are not censored observations but instead represent completely missing survival data. Therefore, we excluded patients with both overall_survival_months and overall_survival_status missing, so that the analysis dataset contains only patients with interpretable OS information for time-to-event modelling.

```
dim(data)          # number of rows and columns

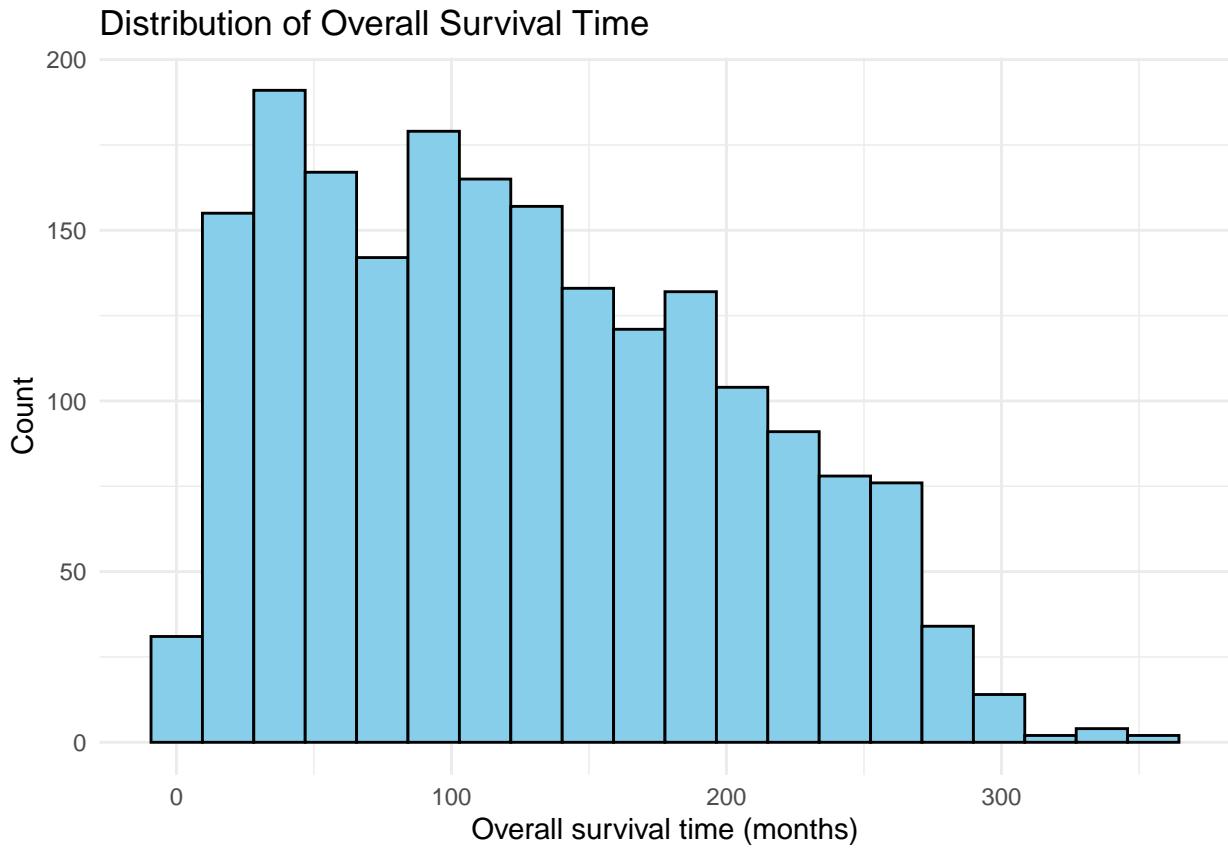
## [1] 1978   34

colnames(data)    # check all variable names

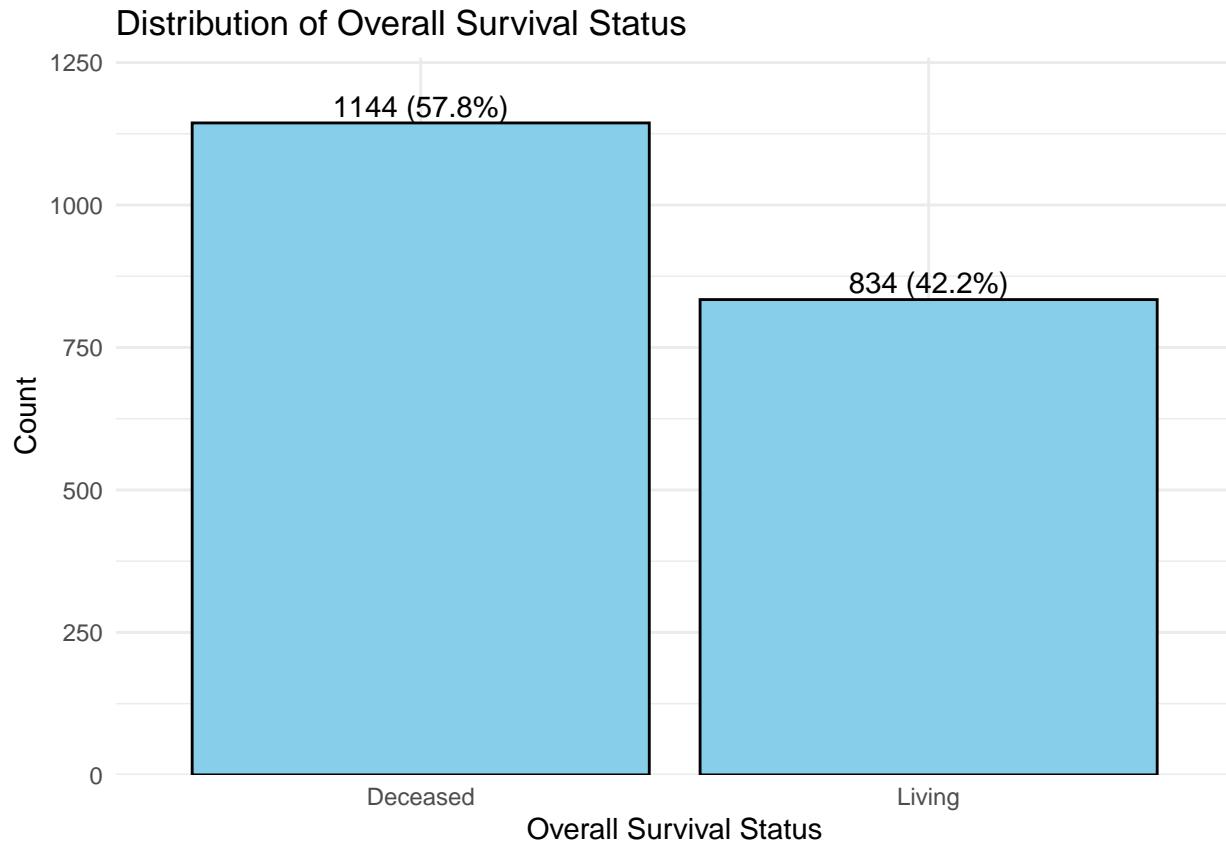
##  [1] "patient_id"           "age_at_diagnosis"
##  [3] "type_of_breast_surgery" "cancer_type"
##  [5] "cancer_type_detailed"   "cellularity"
##  [7] "chemotherapy"          "pam50_claudin_low_subtype"
##  [9] "cohort"                 "er_status_measured_by_ihc"
## [11] "er_status"              "neoplasm_histologic_grade"
## [13] "her2_status_measured_by.snp6" "her2_status"
## [15] "tumor_other_histologic_subtype" "hormone_therapy"
## [17] "inferred_menopausal_state"      "integrative_cluster"
## [19] "primary_tumor_laterality"       "lymph_nodes_examined_positive"
## [21] "mutation_count"                "nottingham_prognostic_index"
## [23] "oncotree_code"                 "overall_survival_months"
## [25] "overall_survival_status"        "pr_status"
## [27] "radio_therapy"                 "relapse_free_status_months"
## [29] "relapse_free_status"           "sex"
## [31] "three_gene_classifier_subtype"  "tumor_size"
## [33] "tumor_stage"                  "vital_status"
```

Visualization of Overall Survival (OS)

```
ggplot(data, aes(x = overall_survival_months)) +
  geom_histogram(bins = 20, color = "black", fill = "skyblue") +
  labs(
    title = "Distribution of Overall Survival Time",
    x = "Overall survival time (months)",
    y = "Count"
  ) +
  theme_minimal()
```



```
ggplot(data, aes(x = factor(overall_survival_status))) +
  geom_bar(aes(y = ..count..), fill = "skyblue", color = "black") + # Count bars
  geom_text(
    stat = "count",
    aes(label = paste0(..count.., " (", round(..count.. / sum(..count..) * 100, 1), "%)")),
    vjust = -0.3,      # move labels slightly above the bar
    color = "black",
    size = 4
  ) +
  scale_y_continuous(
    expand = expansion(mult = c(0, 0.1)) # add 10% headroom on top so labels are fully visible
  ) +
  labs(
    title = "Distribution of Overall Survival Status",
    x = "Overall Survival Status",
    y = "Count"
  ) +
  theme_minimal()
```



Check for missing values:

```
# Proportion and count of missing values per variable (only variables with missing values)
missing_summary <- data %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>% # count missing values
  pivot_longer(
    cols = everything(),
    names_to = "variable",
    values_to = "missing_count"
  ) %>%
  mutate(
    prop_missing = missing_count / nrow(data) # calculate proportion of missing values
  ) %>% # keep only variables with any missing values
  arrange(desc(missing_count)) # sort by missing count (descending)

missing_summary

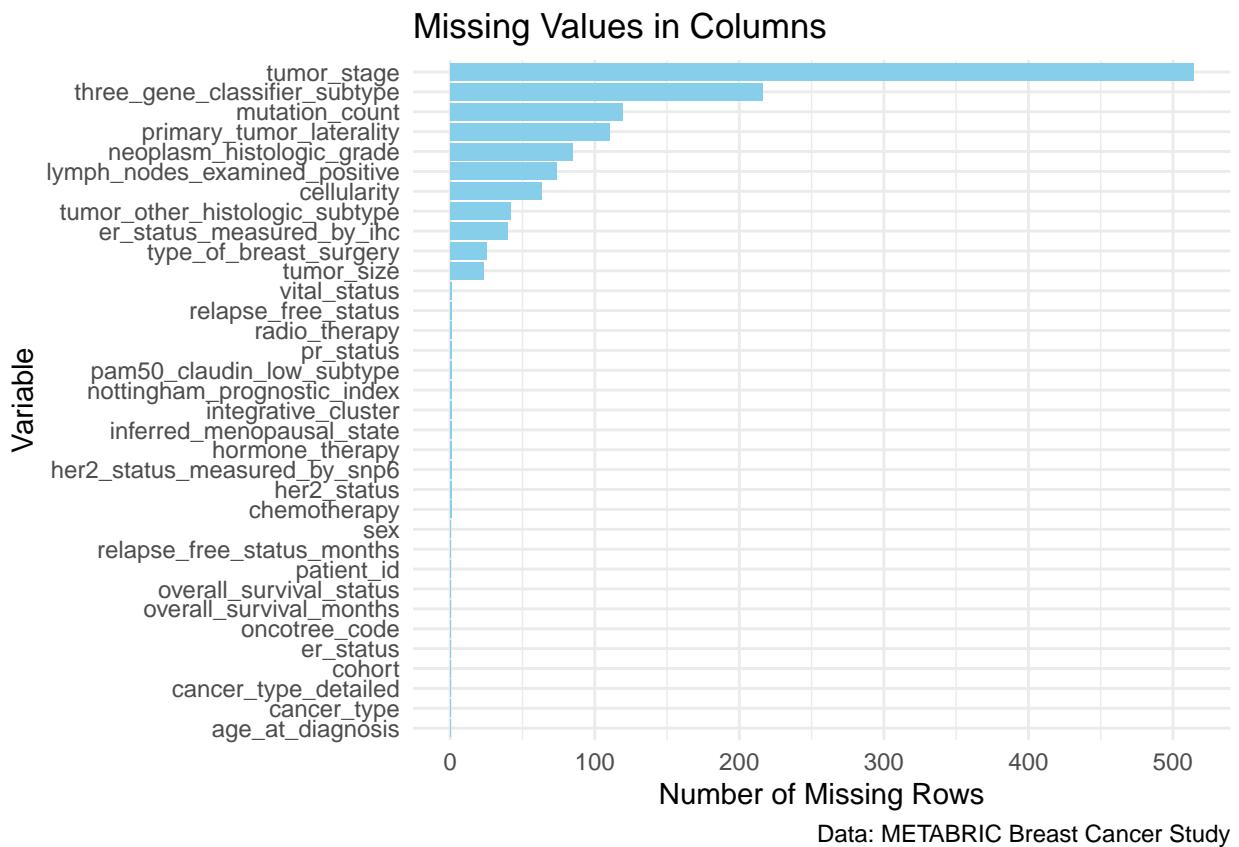
## # A tibble: 34 x 3
##   variable           missing_count  prop_missing
##   <chr>                  <int>        <dbl>
## 1 tumor_stage                514        0.260
## 2 three_gene_classifier_subtype 216        0.109
## 3 mutation_count              119        0.0602
## 4 primary_tumor_laterality    110        0.0556
```

```

## 5 neoplasm_histologic_grade 85 0.0430
## 6 lymph_nodes_examined_positive 74 0.0374
## 7 cellularity 63 0.0319
## 8 tumor_other_histologic_subtype 42 0.0212
## 9 er_status_measured_by_ihc 40 0.0202
## 10 type_of_breast_surgery 25 0.0126
## # i 24 more rows

# Plot missing values per variable (count)
ggplot(missing_summary, aes(x = reorder(variable, missing_count), y = missing_count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Missing Values in Columns",
    x = "Variable",
    y = "Number of Missing Rows",
    caption = "Data: METABRIC Breast Cancer Study"
  ) +
  theme_minimal() +
  coord_flip() # Flip the chart for better readability

```



Define continuous and categorical variables manually:

```

continuous_vars <- c(
  "age_at_diagnosis",
  "tumor_size",
  "lymph_nodes_examined_positive",
  "mutation_count",
  "nottingham_prognostic_index",
  "overall_survival_months",
  "relapse_free_status_months"
)

continuous_vars

```

Continuous variables:

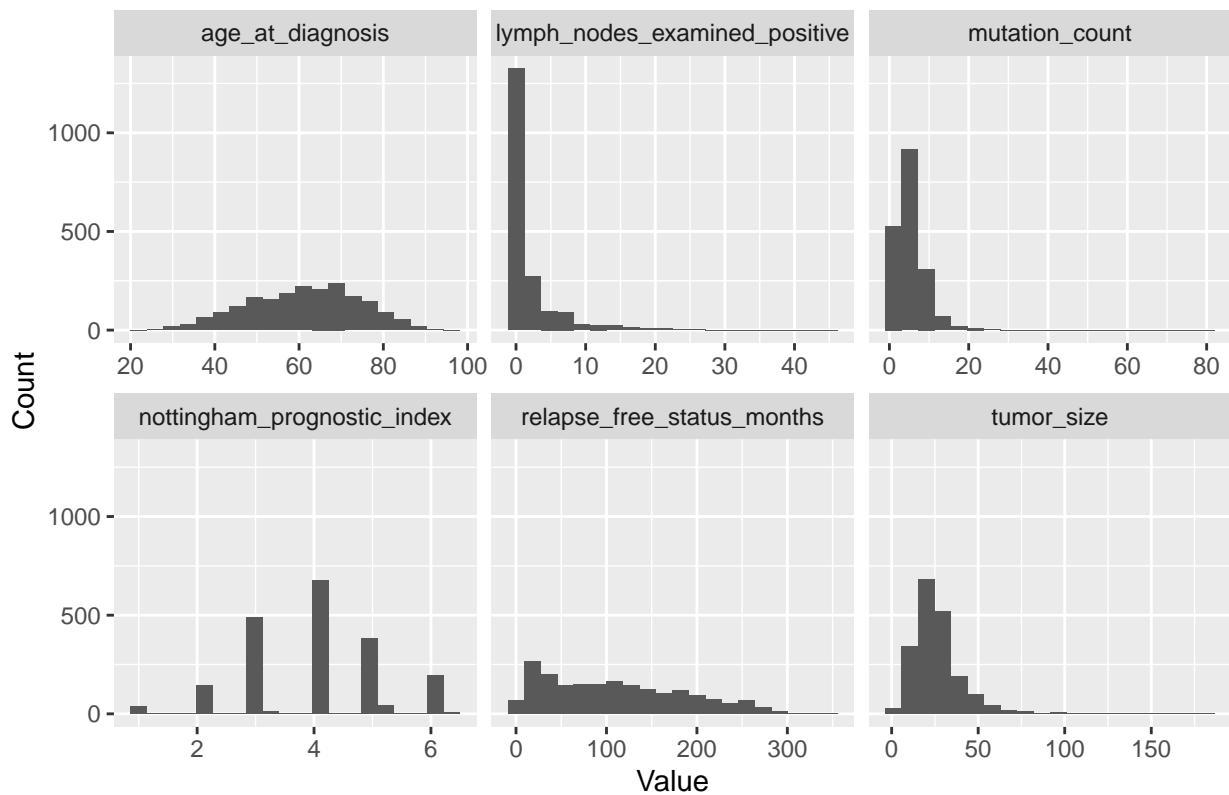
```

## [1] "age_at_diagnosis"           "tumor_size"
## [3] "lymph_nodes_examined_positive" "mutation_count"
## [5] "nottingham_prognostic_index"   "overall_survival_months"
## [7] "relapse_free_status_months"

# Histograms for all continuous variables
data %>%
  select(all_of(setdiff(continuous_vars, "overall_survival_months"))) %>%
  pivot_longer(
    cols = everything(),
    names_to = "variable",
    values_to = "value"
  ) %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 20) +
  facet_wrap(~ variable, scales = "free_x") +
  labs(
    title = "Distributions of continuous variables",
    x = "Value",
    y = "Count"
  )

```

Distributions of continuous variables



```
# Treat other variables as categorical for EDA, but exclude patient_id
categorical_vars <- setdiff(names(data), c(continuous_vars, "patient_id"))
```

```
# Frequency and proportion for each categorical variable (excluding overall_survival_status)
for (var in setdiff(categorical_vars, "overall_survival_status")) {
  cat("\n\n=====\n")
  cat("Variable:", var, "\n")

  tmp <- data %>%
    count(.data[[var]]) %>%
    mutate(prop = n / sum(n)) %>%
    arrange(desc(n))

  print(tmp) # print all rows for this variable
}
```

Categorical variables:

```
##  
##  
## =====
```

```

## Variable: type_of_breast_surgery
##   type_of_breast_surgery     n      prop
## 1           Mastectomy 1169 0.59100101
## 2       Breast Conserving  784 0.39635996
## 3             <NA>    25 0.01263903
##
##
## =====
## Variable: cancer_type
##   cancer_type     n prop
## 1 Breast Cancer 1978    1
##
##
## =====
## Variable: cancer_type_detailed
##   cancer_type_detailed     n      prop
## 1       Breast Invasive Ductal Carcinoma 1537 0.777047523
## 2 Breast Mixed Ductal and Lobular Carcinoma  211 0.106673407
## 3       Breast Invasive Lobular Carcinoma  146 0.073811931
## 4           Invasive Breast Carcinoma    42 0.021233569
## 5   Breast Invasive Mixed Mucinous Carcinoma  23 0.011627907
## 6           Breast                17 0.008594540
## 7       Metaplastic Breast Cancer        2 0.001011122
##
##
## =====
## Variable: cellularity
##   cellularity     n      prop
## 1       High 964 0.48736097
## 2   Moderate 736 0.37209302
## 3       Low 215 0.10869565
## 4         <NA>  63 0.03185035
##
##
## =====
## Variable: chemotherapy
##   chemotherapy     n      prop
## 1          No 1565 0.7912032356
## 2          Yes  412 0.2082912032
## 3         <NA>    1 0.0005055612
##
##
## =====
## Variable: pam50_claudin_low_subtype
##   pam50_claudin_low_subtype     n      prop
## 1                 LumA  700 0.3538928210
## 2                 LumB  475 0.2401415571
## 3                 Her2  224 0.1132457027
## 4       claudin-low 215 0.1086956522
## 5                 Basal 209 0.1056622851
## 6                 Normal 148 0.0748230536
## 7                   NC   6 0.0030333670
## 8         <NA>    1 0.0005055612
##

```

```

## 
## =====
## Variable: cohort
##   cohort    n      prop
## 1       3 763 0.3857432
## 2       1 519 0.2623862
## 3       2 288 0.1456016
## 4       4 238 0.1203236
## 5       5 170 0.0859454
##
##
## =====
## Variable: er_status_measured_by_ihc
##   er_status_measured_by_ihc    n      prop
## 1                         Positve 1499 0.75783620
## 2                         Negative 439 0.22194135
## 3                         <NA>    40 0.02022245
##
##
## =====
## Variable: er_status
##   er_status    n      prop
## 1 Positive 1506 0.7613751
## 2 Negative  472 0.2386249
##
##
## =====
## Variable: neoplasm_histologic_grade
##   neoplasm_histologic_grade    n      prop
## 1                         3 953 0.48179980
## 2                         2 771 0.38978766
## 3                         1 169 0.08543984
## 4                         NA 85 0.04297270
##
##
## =====
## Variable: her2_status_measured_by.snp6
##   her2_status_measured_by.snp6    n      prop
## 1                     Neutral 1433 0.7244691608
## 2                     Gain   438 0.2214357937
## 3                     Loss   101 0.0510616785
## 4                     Undef   5 0.0025278059
## 5                     <NA>   1 0.0005055612
##
##
## =====
## Variable: her2_status
##   her2_status    n      prop
## 1 Negative 1730 0.8746208291
## 2 Positive  247 0.1248736097
## 3 <NA>        1 0.0005055612
##
##
## =====

```

```

## Variable: tumor_other_histologic_subtype
##   tumor_other_histologic_subtype   n      prop
## 1                               Ductal/NST 1491 0.753791709
## 2                               Mixed    211 0.106673407
## 3                               Lobular  146 0.073811931
## 4                               <NA>     42 0.021233569
## 5                               Medullary 25 0.012639029
## 6                               Mucinous 23 0.011627907
## 7       Tubular/ cribriform 21 0.010616785
## 8                               Other    17 0.008594540
## 9                               Metaplastic 2 0.001011122
##
##
## =====
## Variable: hormone_therapy
##   hormone_therapy   n      prop
## 1           Yes 1216 0.6147623862
## 2            No  761 0.3847320526
## 3        <NA>    1 0.0005055612
##
##
## =====
## Variable: inferred_menopausal_state
##   inferred_menopausal_state   n      prop
## 1             Post 1553 0.7851365015
## 2             Pre  424 0.2143579373
## 3        <NA>    1 0.0005055612
##
##
## =====
## Variable: integrative_cluster
##   integrative_cluster   n      prop
## 1                 8 299 0.1511627907
## 2                 3 290 0.1466127401
## 3                4ER+ 259 0.1309403438
## 4                 10 225 0.1137512639
## 5                  5 190 0.0960566229
## 6                  7 190 0.0960566229
## 7                  9 146 0.0738119312
## 8                 1 139 0.0702730030
## 9                  6 85 0.0429726997
## 10                4ER-  82 0.0414560162
## 11                 2 72 0.0364004044
## 12        <NA>    1 0.0005055612
##
##
## =====
## Variable: primary_tumor_laterality
##   primary_tumor_laterality   n      prop
## 1             Left  971 0.49089990
## 2            Right 897 0.45348837
## 3        <NA>  110 0.05561173
##
##

```

```

## =====
## Variable: oncotree_code
##   oncotree_code    n      prop
## 1          IDC 1537 0.777047523
## 2          MDLC 211 0.106673407
## 3          ILC 146 0.073811931
## 4          BRCA 42 0.021233569
## 5          IMMC 23 0.011627907
## 6          BREAST 17 0.008594540
## 7          MBC 2 0.001011122
##
##
## =====
## Variable: pr_status
##   pr_status    n      prop
## 1 Positive 1040 0.5257836198
## 2 Negative 937 0.4737108190
## 3 <NA>     1 0.0005055612
##
##
## =====
## Variable: radio_therapy
##   radio_therapy    n      prop
## 1 Yes 1173 0.5930232558
## 2 No 804 0.4064711830
## 3 <NA>     1 0.0005055612
##
##
## =====
## Variable: relapse_free_status
##   relapse_free_status    n      prop
## 1 Not Recurred 1174 0.5935288170
## 2 Recurred 803 0.4059656218
## 3 <NA>     1 0.0005055612
##
##
## =====
## Variable: sex
##   sex    n prop
## 1 Female 1978 1
##
##
## =====
## Variable: three_gene_classifier_subtype
##   three_gene_classifier_subtype    n      prop
## 1 ER+/HER2- Low Prolif 640 0.3235592
## 2 ER+/HER2- High Prolif 617 0.3119312
## 3 ER-/HER2- 307 0.1552073
## 4 <NA> 216 0.1092012
## 5 HER2+ 198 0.1001011
##
##
## =====
## Variable: tumor_stage

```

```

##   tumor_stage    n      prop
## 1              2 825 0.417087968
## 2            NA 514 0.259858443
## 3              1 500 0.252780586
## 4              3 118 0.059656218
## 5              0  11 0.005561173
## 6              4  10 0.005055612
##
##
## =====
## Variable: vital_status
##           vital_status    n      prop
## 1             Living 834 0.4216380182
## 2       Died of Disease 646 0.3265925177
## 3 Died of Other Causes 497 0.2512639029
## 4          <NA>     1 0.0005055612

```

sex and cancer_type have only a single category (Female and Breast Cancer) in this dataset, making them uninformative for the survival analysis model, as they do not provide any variation in the data.

```

# Remove 'sex' and 'cancer_type' from categorical_vars
categorical_vars <- setdiff(categorical_vars, c("sex", "cancer_type"))

# Print updated categorical_vars
categorical_vars

```

```

## [1] "type_of_breast_surgery"           "cancer_type_detailed"
## [3] "cellularity"                   "chemotherapy"
## [5] "pam50_claudin_low_subtype"      "cohort"
## [7] "er_status_measured_by_ihc"      "er_status"
## [9] "neoplasm_histologic_grade"      "her2_status_measured_by.snp6"
## [11] "her2_status"                   "tumor_other_histologic_subtype"
## [13] "hormone_therapy"                "inferred_menopausal_state"
## [15] "integrative_cluster"           "primary_tumor_laterality"
## [17] "oncotree_code"                 "overall_survival_status"
## [19] "pr_status"                     "radio_therapy"
## [21] "relapse_free_status"           "three_gene_classifier_subtype"
## [23] "tumor_stage"                  "vital_status"

```

Correlation analysis:

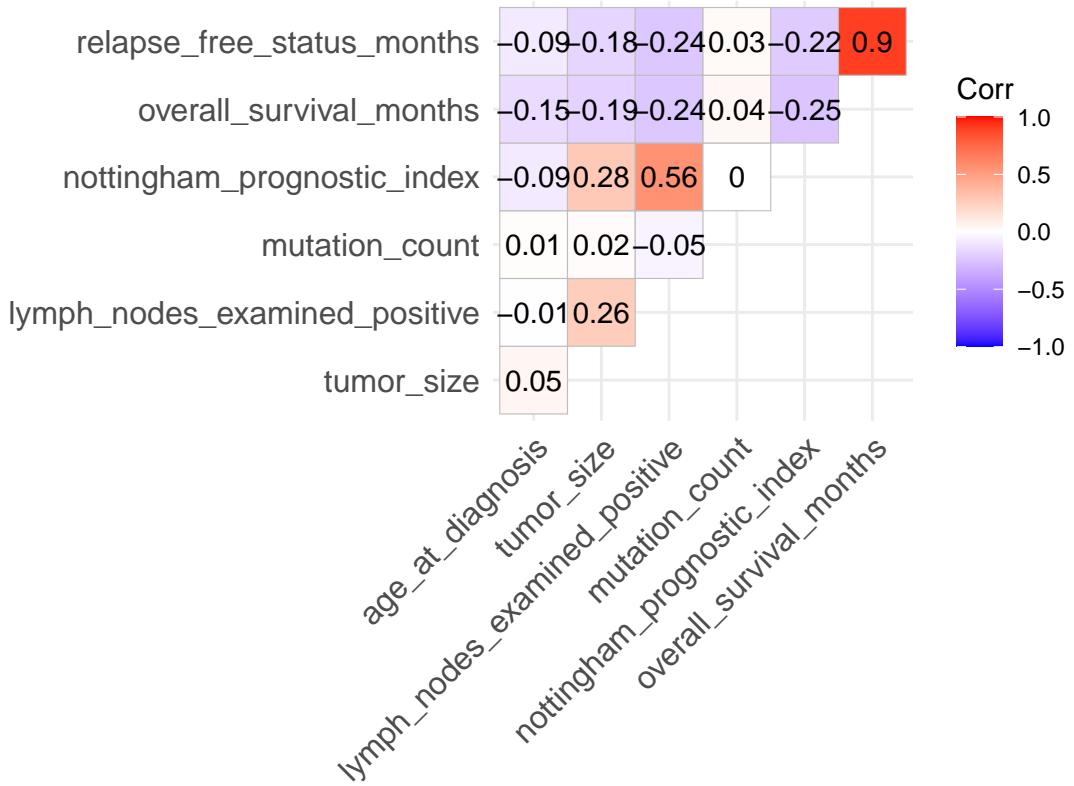
```

# Calculate correlation matrix for continuous variables
cor_matrix <- data %>%
  select(all_of(continuous_vars)) %>%
  cor(use = "pairwise.complete.obs") # handle NA values by pairwise complete observation

ggcorrplot(cor_matrix,
  type = "upper",           # Show only the lower triangle of the matrix
  lab = TRUE,                # Display correlation values
  colors = c("blue", "white", "red"), # Color gradient
  title = "Correlation Matrix of Continuous Variables")

```

Correlation Matrix of Continuous Variables



Based on the correlation matrix shown, `relapse_free_status_months` and `overall_survival_months` exhibit a very strong positive correlation of 0.9, indicating that these two variables are highly related. However, other continuous variables, such as `nottingham_prognostic_index`, `mutation_count`, `lymph_nodes_examined_positive`, and `tumor_size`, do not show high correlations with each other.

If we are conducting survival analysis with `overall_survival_status` as the event indicator, we should exclude `relapse_free_status_months` as a predictor in the model, as it is strongly correlated with `overall_survival_months`. Including both would lead to multicollinearity, where these two variables would essentially provide redundant information, making it difficult to assess their independent effects on survival. Therefore, we should choose `overall_survival_months` as the time variable, as it directly represents the total survival time and includes all relevant information related to the occurrence of the event.

```
# Remove 'relapse_free_status_months' from continuous_vars
continuous_vars <- setdiff(continuous_vars, "relapse_free_status_months")

# Create an empty data frame to store results
chi_results <- data.frame(
  variable = character(),
  p_value = numeric(),
  stringsAsFactors = FALSE
)

# Perform Chi-Square Test for each categorical variable with overall_survival_status
for (var in categorical_vars) {

  # Create a contingency table for each variable and survival status
  contingency_table <- table(data[[var]], data$overall_survival_status)
```

```

# Perform Chi-Square test
chi_test <- chisq.test(contingency_table)

# Store the results in chi_results dataframe
chi_results <- rbind(chi_results, data.frame(variable = var, p_value = chi_test$p.value))

# Print the contingency table and the p-value
cat("\n\n=====\n")
cat("Variable:", var, "\n")
print(contingency_table)
cat("P-value for Chi-Square Test:", chi_test$p.value, "\n")
}

##

## =====
## Variable: type_of_breast_surgery
##
##          Deceased Living
## Breast Conserving      368    416
## Mastectomy              759    410
## P-value for Chi-Square Test: 4.464438e-15

##

## =====
## Variable: cancer_type_detailed
##
##          Deceased Living
## Breast                      5     12
## Breast Invasive Ductal Carcinoma    891    646
## Breast Invasive Lobular Carcinoma    86     60
## Breast Invasive Mixed Mucinous Carcinoma 10     13
## Breast Mixed Ductal and Lobular Carcinoma 132    79
## Invasive Breast Carcinoma           18     24
## Metaplastic Breast Cancer            2      0
## P-value for Chi-Square Test: 0.02097761

##

## =====
## Variable: cellularity
##
##          Deceased Living
## High             562    402
## Low              116     99
## Moderate         436    300
## P-value for Chi-Square Test: 0.3822673

##

## =====
## Variable: chemotherapy
##
##          Deceased Living

```

```

##      No      920     645
##      Yes     223     189
## P-value for Chi-Square Test: 0.09937019

##
##
## =====
## Variable: pam50_claudin_low_subtype
##
##          Deceased Living
##      Basal        115     94
##  claudin-low      94    121
##   Her2         157     67
##   LumA        378    322
##   LumB        315    160
##    NC          5      1
##   Normal       79     69
## P-value for Chi-Square Test: 1.139447e-09
##
##
## =====
## Variable: cohort
##
##          Deceased Living
##    1        227    292
##    2        149    139
##    3        529    234
##    4        124    114
##    5        115     55
## P-value for Chi-Square Test: 2.669872e-20
##
##
## =====
## Variable: er_status_measured_by_ihc
##
##          Deceased Living
##  Negative       246    193
##  Positve       880    619
## P-value for Chi-Square Test: 0.3462201
##
##
## =====
## Variable: er_status
##
##          Deceased Living
##  Negative       262    210
##  Positive       882    624
## P-value for Chi-Square Test: 0.262618
##
##
## =====
## Variable: neoplasm_histologic_grade
##
##          Deceased Living

```

```

##    1      76     93
##    2     432    339
##    3     579    374
## P-value for Chi-Square Test: 0.0003979277

##
##
## =====
## Variable: her2_status_measured_by_snp6
##
##          Deceased Living
##    Gain        270    168
##    Loss         52     49
##    Neutral     818    615
##    Undef        3      2
## P-value for Chi-Square Test: 0.202126
##
##
## =====
## Variable: her2_status
##
##          Deceased Living
##    Negative    988    742
##    Positive     155     92
## P-value for Chi-Square Test: 0.1071625

##
##
## =====
## Variable: tumor_other_histologic_subtype
##
##          Deceased Living
##    Ductal/NST     871    620
##    Lobular       86     60
##    Medullary      14     11
##    Metaplastic      2      0
##    Mixed         132    79
##    Mucinous       10     13
##    Other          5     12
##    Tubular/ cribriform   6     15
## P-value for Chi-Square Test: 0.009518724
##
##
## =====
## Variable: hormone_therapy
##
##          Deceased Living
##    No        426    335
##    Yes       717    499
## P-value for Chi-Square Test: 0.2073769
##
##
## =====
## Variable: inferred_menopausal_state

```

```

##
##          Deceased Living
##    Post      966     587
##    Pre       177     247
## P-value for Chi-Square Test: 6.178334e-14
##
##
## =====
## Variable: integrative_cluster
##
##          Deceased Living
##    1        80      59
##   10       107     118
##    2        52      20
##    3       151     139
##   4ER-      44      38
##   4ER+     130     129
##    5       130      60
##    6        59      26
##    7       113      77
##    8       179     120
##    9        98      48
## P-value for Chi-Square Test: 1.504806e-06
##
##
## =====
## Variable: primary_tumor_laterality
##
##          Deceased Living
##    Left      571     400
##   Right     495     402
## P-value for Chi-Square Test: 0.1252675

##
##
## =====
## Variable: oncotree_code
##
##          Deceased Living
##    BRCA      18      24
##   BREAST     5      12
##    IDC      891     646
##   ILC       86      60
##   IMMC      10      13
##   MBC        2      0
##   MDLC     132      79
## P-value for Chi-Square Test: 0.02097761
##
##
## =====
## Variable: overall_survival_status
##
##          Deceased Living
##    Deceased   1144      0

```

```

##   Living          0    834
## P-value for Chi-Square Test: 0
##
## 
## -----
## Variable: pr_status
##
##           Deceased Living
##   Negative      551    386
##   Positive      592    448
## P-value for Chi-Square Test: 0.4235425
##
## 
## -----
## Variable: radio_therapy
##
##           Deceased Living
##   No        514    290
##   Yes       629    544
## P-value for Chi-Square Test: 6.420301e-06
##
## 
## -----
## Variable: relapse_free_status
##
##           Deceased Living
##   Not Recurred   432    742
##   Recurred       711     92
## P-value for Chi-Square Test: 2.106947e-115
##
## 
## -----
## Variable: three_gene_classifier_subtype
##
##           Deceased Living
##   ER-/HER2-        153    154
##   ER+/HER2- High Prolif   400    217
##   ER+/HER2- Low Prolif   329    311
##   HER2+            121     77
## P-value for Chi-Square Test: 7.079376e-07

##
## 
## -----
## Variable: tumor_stage
##
##           Deceased Living
##   0            2     9
##   1          228   272
##   2          497   328
##   3           87    31
##   4            9     1
## P-value for Chi-Square Test: 4.935459e-11
##

```

```

## 
## =====
## Variable: vital_status
##
##          Deceased Living
## Died of Disease      646     0
## Died of Other Causes 497     0
## Living                  0    834
## P-value for Chi-Square Test: 0

# Print the final results with p-values for all variables except overall_survival_status
chi_results

```

```

##           variable      p_value
## 1 type_of_breast_surgery 4.464438e-15
## 2 cancer_type_detailed 2.097761e-02
## 3 cellularity          3.822673e-01
## 4 chemotherapy          9.937019e-02
## 5 pam50_claudin_low_subtype 1.139447e-09
## 6 cohort                 2.669872e-20
## 7 er_status_measured_by_ihc 3.462201e-01
## 8 er_status              2.626180e-01
## 9 neoplasm_histologic_grade 3.979277e-04
## 10 her2_status_measured_by_snp6 2.021260e-01
## 11 her2_status            1.071625e-01
## 12 tumor_other_histologic_subtype 9.518724e-03
## 13 hormone_therapy        2.073769e-01
## 14 inferred_menopausal_state 6.178334e-14
## 15 integrative_cluster    1.504806e-06
## 16 primary_tumor_laterality 1.252675e-01
## 17 oncotree_code           2.097761e-02
## 18 overall_survival_status 0.000000e+00
## 19 pr_status               4.235425e-01
## 20 radio_therapy            6.420301e-06
## 21 relapse_free_status     2.106947e-115
## 22 three_gene_classifier_subtype 7.079376e-07
## 23 tumor_stage              4.935459e-11
## 24 vital_status             0.000000e+00

```

Since “vital_status” variable is already directly related to the outcome of survival and is a proxy for the status of the patient, it is unnecessary to include it in the model. This variable is highly correlated with survival, and adding it could lead to perfect prediction of the outcome or multicollinearity.

Although the chi-square test showed a statistically significant association between “cohort” and overall survival status, we decided not to include “cohort” as a covariate in the survival model. In this dataset, “cohort” primarily reflects study batch or recruitment group rather than a meaningful clinical or biological characteristic of the patients. Therefore, it is better interpreted as a design-related or administrative label, rather than a true risk factor for overall survival.

Although “relapse_free_status” shows a highly significant association with overall survival status in the chi-square test, we chose not to include it as a covariate in the overall survival model. Clinically, relapse is an intermediate outcome that occurs after baseline and lies on the causal pathway between baseline risk factors and death. Treating relapse status as a predictor of overall survival would therefore introduce information leakage and could distort the estimated effects of true baseline covariates. In addition, “relapse_free_status”

is closely related to “relapse_free_status_months”, which we already excluded as an alternative time-to-event endpoint.

```
# Filter results to only show variables with p-value < 0.05 and exclude 'vital_status' and 'cohort'
significant_vars <- chi_results %>%
  filter(p_value < 0.05) %>%
  filter(!variable %in% c("vital_status", "cohort", "relapse_free_status"))

# Print the significant variables
significant_vars
```

	variable	p_value
## 1	type_of_breast_surgery	4.464438e-15
## 2	cancer_type_detailed	2.097761e-02
## 3	pam50_claudin_low_subtype	1.139447e-09
## 4	neoplasm_histologic_grade	3.979277e-04
## 5	tumor_other_histologic_subtype	9.518724e-03
## 6	inferred_menopausal_state	6.178334e-14
## 7	integrative_cluster	1.504806e-06
## 8	oncotree_code	2.097761e-02
## 9	overall_survival_status	0.000000e+00
## 10	radio_therapy	6.420301e-06
## 11	three_gene_classifier_subtype	7.079376e-07
## 12	tumor_stage	4.935459e-11

Final dataset after variable selection

```
# Extract significant variables' names from significant_vars (filter for p-value < 0.05)
significant_var_names <- significant_vars$variable

# Filter the data by including only the selected variables from both significant_vars and continuous_vars
selected_vars <- c(significant_var_names, continuous_vars)

# Create a new dataset with only the selected variables
filtered_data <- data %>%
  select(all_of(selected_vars))

# View the filtered dataset
head(filtered_data)
```

	type_of_breast_surgery	cancer_type_detailed
## 1	Mastectomy	Breast Invasive Ductal Carcinoma
## 2	Breast Conserving	Breast Invasive Ductal Carcinoma
## 3	Mastectomy	Breast Invasive Ductal Carcinoma
## 4	Mastectomy	Breast Mixed Ductal and Lobular Carcinoma
## 5	Mastectomy	Breast Mixed Ductal and Lobular Carcinoma
## 6	Mastectomy	Breast Invasive Ductal Carcinoma
## 1	pam50_claudin_low_subtype	neoplasm_histologic_grade
## 2	claudin-low	3
## 3	LumA	3
	LumB	2

```

## 4 LumB 2
## 5 LumB 3
## 6 LumB 3
## tumor_other_histologic_subtype inferred_menopausal_state integrative_cluster
## 1 Ductal/NST Post 4ER+
## 2 Ductal/NST Pre 4ER+
## 3 Ductal/NST Pre 3
## 4 Mixed Pre 9
## 5 Mixed Post 9
## 6 Ductal/NST Post 7
## oncotree_code overall_survival_status radio_therapy
## 1 IDC Living Yes
## 2 IDC Living Yes
## 3 IDC Deceased No
## 4 MDLC Living Yes
## 5 MDLC Deceased Yes
## 6 IDC Deceased Yes
## three_gene_classifier_subtype tumor_stage age_at_diagnosis tumor_size
## 1 ER-/HER2- 2 75.65 22
## 2 ER+/HER2- High Prolif 1 43.19 10
## 3 <NA> 2 48.87 15
## 4 <NA> 2 47.68 25
## 5 ER+/HER2- High Prolif 2 76.97 40
## 6 ER+/HER2- High Prolif 4 78.77 31
## lymph_nodes_examined_positive mutation_count nottingham_prognostic_index
## 1 10 NA 6.044
## 2 0 2 4.020
## 3 1 2 4.030
## 4 3 1 4.050
## 5 8 2 6.080
## 6 0 4 4.062
## overall_survival_months
## 1 140.50000
## 2 84.63333
## 3 163.70000
## 4 164.93333
## 5 41.36667
## 6 7.80000

```

Overview of demographic and baseline variables (Table 1)

```

filtered_data <- filtered_data |>
  mutate(overall_survival_status = factor(overall_survival_status))

vars_for_table1 <- c(
  "age_at_diagnosis",
  "tumor_size",
  "lymph_nodes_examined_positive",
  "mutation_count",
  "nottingham_prognostic_index",
  "type_of_breast_surgery",
  "cancer_type_detailed",

```

```

"pam50_claudin_low_subtype",
"neoplasm_histologic_grade",
"tumor_other_histologic_subtype",
"inferred_menopausal_state",
"integrative_cluster",
"radio_therapy",
"three_gene_classifier_subtype",
"tumor_stage"
)

continuous_vars <- c(
  "age_at_diagnosis",
  "tumor_size",
  "lymph_nodes_examined_positive",
  "mutation_count",
  "nottingham_prognostic_index"
)
categorical_vars <- setdiff(vars_for_table1, continuous_vars)

pvalue_fmt <- function(x) {
  ifelse(x < 0.001, "<0.001", formatC(x, format = "f", digits = 3))
}

table1 <-
  filtered_data |>
  select(all_of(c("overall_survival_status", vars_for_table1))) |>
 tbl_summary(
    by = overall_survival_status,
    type = list(
      all_of(continuous_vars) ~ "continuous2",
      all_of(categorical_vars) ~ "categorical"
    ),
    statistic = list(
      all_of(continuous_vars) ~ "{mean} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    missing = "ifany"
  ) |>
  add_n() |>
  add_p(
    test = list(
      all_of(continuous_vars) ~ "t.test",
      all_categorical() ~ "chisq.test"
    ),
    pvalue_fun = pvalue_fmt
  ) |>
  modify_caption("**Table 1. Baseline characteristics by overall survival status**") |>
  bold_labels()

```

```

## The following warnings were returned during 'modify_caption()':
## ! For variable 'cancer_type_detailed' ('overall_survival_status') and
##   "statistic", "p.value", and "parameter" statistics: Chi-squared approximation
##   may be incorrect

```

```

## ! For variable 'pam50_claudin_low_subtype' ('overall_survival_status') and
##   "statistic", "p.value", and "parameter" statistics: Chi-squared approximation
##   may be incorrect
## ! For variable 'tumor_other_histologic_subtype' ('overall_survival_status') and
##   "statistic", "p.value", and "parameter" statistics: Chi-squared approximation
##   may be incorrect
## ! For variable 'tumor_stage' ('overall_survival_status') and "statistic",
##   "p.value", and "parameter" statistics: Chi-squared approximation may be
##   incorrect

```

```
table1
```

Multivariable Cox Model

```

library(survival)
library(broom)

final_inter_mod <- coxph(
  Surv(
    overall_survival_months,
    overall_survival_status == "Deceased"    # event indicator
  ) ~
  age_at_diagnosis +
  tumor_size +
  lymph_nodes_examined_positive +
  nottingham_prognostic_index +
  radio_therapy +
  inferred_menopausal_state +
  tumor_stage +
  three_gene_classifier_subtype +
  type_of_breast_surgery +
  cancer_type_detailed +
  pam50_claudin_low_subtype +
  tumor_other_histologic_subtype +
  integrative_cluster +
  mutation_count +
  radio_therapy:tumor_size, # treatment x tumor size interaction
  data = filtered_data
)

```

Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
Loglik converged before variable 12,13,14,15 ; coefficient may be infinite.

```
summary(final_inter_mod)
```

```

## Call:
## coxph(formula = Surv(overall_survival_months, overall_survival_status ==
##   "Deceased") ~ age_at_diagnosis + tumor_size + lymph_nodes_examined_positive +
##   nottingham_prognostic_index + radio_therapy + inferred_menopausal_state +
##   tumor_stage + three_gene_classifier_subtype + type_of_breast_surgery +

```

```

##      cancer_type_detailed + pam50_claudin_low_subtype + tumor_other_histologic_subtype +
##      integrative_cluster + mutation_count + radio_therapy:tumor_size,
##      data = filtered_data)
##
##      n= 1186, number of events= 665
##      (792 observations deleted due to missingness)
##
##                                     coef
## age_at_diagnosis                  4.915e-02
## tumor_size                         1.362e-02
## lymph_nodes_examined_positive      5.022e-02
## nottingham_prognostic_index       1.176e-01
## radio_therapyYes                   -2.313e-02
## inferred_menopausal_statePre     5.351e-01
## tumor_stage                        8.522e-02
## three_gene_classifier_subtypeER+/HER2- High Prolif -1.273e-01
## three_gene_classifier_subtypeER+/HER2- Low Prolif -2.845e-01
## three_gene_classifier_subtypeHER2+ -4.093e-01
## type_of_breast_surgeryMastectomy  1.047e-01
## cancer_type_detailedBreast Invasive Ductal Carcinoma 1.518e+01
## cancer_type_detailedBreast Invasive Lobular Carcinoma 1.495e+01
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma 1.454e+01
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma 1.521e+01
## pam50_claudin_low_subtypeclaudin-low -2.838e-01
## pam50_claudin_low_subtypeHer2       8.155e-02
## pam50_claudin_low_subtypeLumA      -2.305e-01
## pam50_claudin_low_subtypeLumB      -2.020e-01
## pam50_claudin_low_subtypeNC        -9.304e-01
## pam50_claudin_low_subtypeNormal   1.491e-01
## tumor_other_histologic_subtypeLobular NA
## tumor_other_histologic_subtypeMedullary 3.351e-01
## tumor_other_histologic_subtypeMixed NA
## tumor_other_histologic_subtypeMucinous NA
## tumor_other_histologic_subtypeOther NA
## tumor_other_histologic_subtypeTubular/ cribriform -5.608e-01
## integrative_cluster10             -3.848e-01
## integrative_cluster2              3.289e-02
## integrative_cluster3              1.285e-02
## integrative_cluster4ER-          -1.381e-01
## integrative_cluster4ER+          -1.361e-01
## integrative_cluster5              5.965e-01
## integrative_cluster6              -7.270e-02
## integrative_cluster7              -7.710e-02
## integrative_cluster8              -4.029e-02
## integrative_cluster9              -6.470e-02
## mutation_count                    4.164e-03
## tumor_size:radio_therapyYes       -6.761e-03
##                                     exp(coef)
## age_at_diagnosis                  1.050e+00
## tumor_size                         1.014e+00
## lymph_nodes_examined_positive      1.052e+00
## nottingham_prognostic_index       1.125e+00
## radio_therapyYes                   9.771e-01
## inferred_menopausal_statePre     1.708e+00

```

## tumor_stage		1.089e+00
## three_gene_classifier_subtypeER+/HER2- High Prolif		8.804e-01
## three_gene_classifier_subtypeER+/HER2- Low Prolif		7.524e-01
## three_gene_classifier_subtypeHER2+		6.641e-01
## type_of_breast_surgeryMastectomy		1.110e+00
## cancer_type_detailedBreast Invasive Ductal Carcinoma		3.911e+06
## cancer_type_detailedBreast Invasive Lobular Carcinoma		3.111e+06
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma		2.067e+06
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma		4.013e+06
## pam50_claudin_low_subtypeclaudin-low		7.529e-01
## pam50_claudin_low_subtypeHer2		1.085e+00
## pam50_claudin_low_subtypeLumA		7.941e-01
## pam50_claudin_low_subtypeLumB		8.171e-01
## pam50_claudin_low_subtypeNC		3.944e-01
## pam50_claudin_low_subtypeNormal		1.161e+00
## tumor_other_histologic_subtypeLobular		NA
## tumor_other_histologic_subtypeMedullary		1.398e+00
## tumor_other_histologic_subtypeMixed		NA
## tumor_other_histologic_subtypeMucinous		NA
## tumor_other_histologic_subtypeOther		NA
## tumor_other_histologic_subtypeTubular/ cribriform		5.708e-01
## integrative_cluster10		6.806e-01
## integrative_cluster2		1.033e+00
## integrative_cluster3		1.013e+00
## integrative_cluster4ER-		8.710e-01
## integrative_cluster4ER+		8.727e-01
## integrative_cluster5		1.816e+00
## integrative_cluster6		9.299e-01
## integrative_cluster7		9.258e-01
## integrative_cluster8		9.605e-01
## integrative_cluster9		9.373e-01
## mutation_count		1.004e+00
## tumor_size:radio_therapyYes		9.933e-01
## age_at_diagnosis	se(coef)	z
## tumor_size	4.881e-03	10.069
## lymph_nodes_examined_positive	5.780e-03	2.355
## nottingham_prognostic_index	1.188e-02	4.229
## radio_therapyYes	5.235e-02	2.247
## inferred_menopausal_statePre	1.990e-01	-0.116
## tumor_stage	1.569e-01	3.411
## three_gene_classifier_subtypeER+/HER2- High Prolif	8.412e-02	1.013
## three_gene_classifier_subtypeER+/HER2- Low Prolif	2.136e-01	-0.596
## three_gene_classifier_subtypeHER2+	2.191e-01	-1.299
## type_of_breast_surgeryMastectomy	2.918e-01	-1.403
## cancer_type_detailedBreast Invasive Ductal Carcinoma	1.052e-01	0.995
## cancer_type_detailedBreast Invasive Lobular Carcinoma	8.744e+02	0.017
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma	8.744e+02	0.017
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma	8.744e+02	0.017
## pam50_claudin_low_subtypeclaudin-low	2.110e-01	-1.345
## pam50_claudin_low_subtypeHer2	2.221e-01	0.367
## pam50_claudin_low_subtypeLumA	2.330e-01	-0.989
## pam50_claudin_low_subtypeLumB	2.347e-01	-0.861
## pam50_claudin_low_subtypeNC	1.035e+00	-0.899

## pam50_claudin_low_subtypeNormal	2.608e-01	0.572
## tumor_other_histologic_subtypeLobular	0.000e+00	NA
## tumor_other_histologic_subtypeMedullary	3.487e-01	0.961
## tumor_other_histologic_subtypeMixed	0.000e+00	NA
## tumor_other_histologic_subtypeMucinous	0.000e+00	NA
## tumor_other_histologic_subtypeOther	0.000e+00	NA
## tumor_other_histologic_subtypeTubular/ cribriform	4.581e-01	-1.224
## integrative_cluster10	2.459e-01	-1.565
## integrative_cluster2	2.401e-01	0.137
## integrative_cluster3	2.048e-01	0.063
## integrative_cluster4ER-	2.905e-01	-0.476
## integrative_cluster4ER+	2.089e-01	-0.652
## integrative_cluster5	2.901e-01	2.056
## integrative_cluster6	2.331e-01	-0.312
## integrative_cluster7	2.110e-01	-0.365
## integrative_cluster8	1.991e-01	-0.202
## integrative_cluster9	2.139e-01	-0.303
## mutation_count	8.942e-03	0.466
## tumor_size:radio_therapyYes	5.965e-03	-1.133
##	Pr(> z)	
## age_at_diagnosis	< 2e-16 ***	
## tumor_size	0.018504 *	
## lymph_nodes_examined_positive	2.35e-05 ***	
## nottingham_prognostic_index	0.024663 *	
## radio_therapyYes	0.907472	
## inferred_menopausal_statePre	0.000647 ***	
## tumor_stage	0.311063	
## three_gene_classifier_subtypeER+/HER2- High Prolif	0.551151	
## three_gene_classifier_subtypeER+/HER2- Low Prolif	0.194077	
## three_gene_classifier_subtypeHER2+	0.160643	
## type_of_breast_surgeryMastectomy	0.319882	
## cancer_type_detailedBreast Invasive Ductal Carcinoma	0.986149	
## cancer_type_detailedBreast Invasive Lobular Carcinoma	0.986358	
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma	0.986731	
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma	0.986126	
## pam50_claudin_low_subtypeclaudin-low	0.178585	
## pam50_claudin_low_subtypeHer2	0.713564	
## pam50_claudin_low_subtypeLumA	0.322483	
## pam50_claudin_low_subtypeLumB	0.389483	
## pam50_claudin_low_subtypeNC	0.368475	
## pam50_claudin_low_subtypeNormal	0.567480	
## tumor_other_histologic_subtypeLobular	NA	
## tumor_other_histologic_subtypeMedullary	0.336487	
## tumor_other_histologic_subtypeMixed	NA	
## tumor_other_histologic_subtypeMucinous	NA	
## tumor_other_histologic_subtypeOther	NA	
## tumor_other_histologic_subtypeTubular/ cribriform	0.220930	
## integrative_cluster10	0.117660	
## integrative_cluster2	0.891047	
## integrative_cluster3	0.949964	
## integrative_cluster4ER-	0.634429	
## integrative_cluster4ER+	0.514676	
## integrative_cluster5	0.039742 *	
## integrative_cluster6	0.755177	

```

## integrative_cluster7          0.714836
## integrative_cluster8          0.839616
## integrative_cluster9          0.762239
## mutation_count                0.641432
## tumor_size:radio_therapyYes   0.257030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                         exp(coef)
## age_at_diagnosis                  1.050e+00
## tumor_size                        1.014e+00
## lymph_nodes_examined_positive     1.052e+00
## nottingham_prognostic_index      1.125e+00
## radio_therapyYes                 9.771e-01
## inferred_menopausal_statePre    1.708e+00
## tumor_stage                       1.089e+00
## three_gene_classifier_subtypeER+/HER2- High Prolif 8.804e-01
## three_gene_classifier_subtypeER+/HER2- Low Prolif 7.524e-01
## three_gene_classifier_subtypeHER2+ 6.641e-01
## type_of_breast_surgeryMastectomy 1.110e+00
## cancer_type_detailedBreast Invasive Ductal Carcinoma 3.911e+06
## cancer_type_detailedBreast Invasive Lobular Carcinoma 3.111e+06
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma 2.067e+06
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma 4.013e+06
## pam50_claudin_low_subtypeclaudin-low 7.529e-01
## pam50_claudin_low_subtypeHer2     1.085e+00
## pam50_claudin_low_subtypeLumA    7.941e-01
## pam50_claudin_low_subtypeLumB    8.171e-01
## pam50_claudin_low_subtypeNC      3.944e-01
## pam50_claudin_low_subtypeNormal 1.161e+00
## tumor_other_histologic_subtypeLobular NA
## tumor_other_histologic_subtypeMedullary 1.398e+00
## tumor_other_histologic_subtypeMixed NA
## tumor_other_histologic_subtypeMucinous NA
## tumor_other_histologic_subtypeOther NA
## tumor_other_histologic_subtypeTubular/ cribriform 5.708e-01
## integrative_cluster10            6.806e-01
## integrative_cluster2             1.033e+00
## integrative_cluster3             1.013e+00
## integrative_cluster4ER-          8.710e-01
## integrative_cluster4ER+          8.727e-01
## integrative_cluster5             1.816e+00
## integrative_cluster6             9.299e-01
## integrative_cluster7             9.258e-01
## integrative_cluster8             9.605e-01
## integrative_cluster9             9.373e-01
## mutation_count                  1.004e+00
## tumor_size:radio_therapyYes     9.933e-01
##                                         exp(-coef)
## age_at_diagnosis                9.520e-01
## tumor_size                       9.865e-01
## lymph_nodes_examined_positive    9.510e-01
## nottingham_prognostic_index     8.890e-01
## radio_therapyYes                 1.023e+00

```

## inferred_menopausal_statePre	5.856e-01
## tumor_stage	9.183e-01
## three_gene_classifier_subtypeER+/HER2- High Prolif	1.136e+00
## three_gene_classifier_subtypeER+/HER2- Low Prolif	1.329e+00
## three_gene_classifier_subtypeHER2+	1.506e+00
## type_of_breast_surgeryMastectomy	9.006e-01
## cancer_type_detailedBreast Invasive Ductal Carcinoma	2.557e-07
## cancer_type_detailedBreast Invasive Lobular Carcinoma	3.214e-07
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma	4.839e-07
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma	2.492e-07
## pam50_claudin_low_subtypeclaudin-low	1.328e+00
## pam50_claudin_low_subtypeHer2	9.217e-01
## pam50_claudin_low_subtypeLumA	1.259e+00
## pam50_claudin_low_subtypeLumB	1.224e+00
## pam50_claudin_low_subtypeNC	2.535e+00
## pam50_claudin_low_subtypeNormal	8.615e-01
## tumor_other_histologic_subtypeLobular	NA
## tumor_other_histologic_subtypeMedullary	7.153e-01
## tumor_other_histologic_subtypeMixed	NA
## tumor_other_histologic_subtypeMucinous	NA
## tumor_other_histologic_subtypeOther	NA
## tumor_other_histologic_subtypeTubular/ cribriform	1.752e+00
## integrative_cluster10	1.469e+00
## integrative_cluster2	9.676e-01
## integrative_cluster3	9.872e-01
## integrative_cluster4ER-	1.148e+00
## integrative_cluster4ER+	1.146e+00
## integrative_cluster5	5.507e-01
## integrative_cluster6	1.075e+00
## integrative_cluster7	1.080e+00
## integrative_cluster8	1.041e+00
## integrative_cluster9	1.067e+00
## mutation_count	9.958e-01
## tumor_size:radio_therapyYes	1.007e+00
##	lower .95
## age_at_diagnosis	1.04038
## tumor_size	1.00229
## lymph_nodes_examined_positive	1.02731
## nottingham_prognostic_index	1.01512
## radio_therapyYes	0.66153
## inferred_menopausal_statePre	1.25565
## tumor_stage	0.92343
## three_gene_classifier_subtypeER+/HER2- High Prolif	0.57922
## three_gene_classifier_subtypeER+/HER2- Low Prolif	0.48976
## three_gene_classifier_subtypeHER2+	0.37486
## type_of_breast_surgeryMastectomy	0.90340
## cancer_type_detailedBreast Invasive Ductal Carcinoma	0.00000
## cancer_type_detailedBreast Invasive Lobular Carcinoma	0.00000
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma	0.00000
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma	0.00000
## pam50_claudin_low_subtypeclaudin-low	0.49791
## pam50_claudin_low_subtypeHer2	0.70197
## pam50_claudin_low_subtypeLumA	0.50303
## pam50_claudin_low_subtypeLumB	0.51579

## pam50_claudin_low_subtypeNC	0.05192
## pam50_claudin_low_subtypeNormal	0.69625
## tumor_other_histologic_subtypeLobular	NA
## tumor_other_histologic_subtypeMedullary	0.70592
## tumor_other_histologic_subtypeMixed	NA
## tumor_other_histologic_subtypeMucinous	NA
## tumor_other_histologic_subtypeOther	NA
## tumor_other_histologic_subtypeTubular/ cribriform	0.23253
## integrative_cluster10	0.42027
## integrative_cluster2	0.64553
## integrative_cluster3	0.67807
## integrative_cluster4ER-	0.49285
## integrative_cluster4ER+	0.57954
## integrative_cluster5	1.02838
## integrative_cluster6	0.58881
## integrative_cluster7	0.61221
## integrative_cluster8	0.65018
## integrative_cluster9	0.61639
## mutation_count	0.98673
## tumor_size:radio_therapyYes	0.98172
##	upper .95
## age_at_diagnosis	1.060
## tumor_size	1.025
## lymph_nodes_examined_positive	1.076
## nottingham_prognostic_index	1.246
## radio_therapyYes	1.443
## inferred_menopausal_statePre	2.322
## tumor_stage	1.284
## three_gene_classifier_subtypeER+/HER2- High Prolif	1.338
## three_gene_classifier_subtypeER+/HER2- Low Prolif	1.156
## three_gene_classifier_subtypeHER2+	1.176
## type_of_breast_surgeryMastectomy	1.365
## cancer_type_detailedBreast Invasive Ductal Carcinoma	Inf
## cancer_type_detailedBreast Invasive Lobular Carcinoma	Inf
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma	Inf
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma	Inf
## pam50_claudin_low_subtypeclaudin-low	1.139
## pam50_claudin_low_subtypeHer2	1.677
## pam50_claudin_low_subtypeLumA	1.254
## pam50_claudin_low_subtypeLumB	1.294
## pam50_claudin_low_subtypeNC	2.996
## pam50_claudin_low_subtypeNormal	1.935
## tumor_other_histologic_subtypeLobular	NA
## tumor_other_histologic_subtypeMedullary	2.769
## tumor_other_histologic_subtypeMixed	NA
## tumor_other_histologic_subtypeMucinous	NA
## tumor_other_histologic_subtypeOther	NA
## tumor_other_histologic_subtypeTubular/ cribriform	1.401
## integrative_cluster10	1.102
## integrative_cluster2	1.654
## integrative_cluster3	1.513
## integrative_cluster4ER-	1.539
## integrative_cluster4ER+	1.314
## integrative_cluster5	3.206

```

## integrative_cluster6           1.469
## integrative_cluster7          1.400
## integrative_cluster8          1.419
## integrative_cluster9          1.425
## mutation_count                1.022
## tumor_size:radio_therapyYes   1.005
##
## Concordance= 0.693  (se = 0.011 )
## Likelihood ratio test= 330  on 35 df,  p=<2e-16
## Wald test                  = 339.5  on 35 df,  p=<2e-16
## Score (logrank) test = 379  on 35 df,  p=<2e-16

cox_tidy <- broom::tidy(
  final_inter_mod,
  exponentiate = TRUE,
  conf.int     = TRUE
)

cox_tidy

## # A tibble: 39 x 7
##   term            estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>    <dbl>     <dbl>    <dbl>    <dbl>    <dbl>
## 1 age_at_diagnosis  1.05    0.00488   10.1    7.56e-24  1.04    1.06
## 2 tumor_size        1.01    0.00578   2.36    1.85e- 2  1.00    1.03
## 3 lymph_nodes_examined 1.05    0.0119    4.23    2.35e- 5  1.03    1.08
## 4 nottingham_prognostic 1.12    0.0523    2.25    2.47e- 2  1.02    1.25
## 5 radio_therapyYes  0.977   0.199     -0.116   9.07e- 1  0.662   1.44
## 6 inferred_menopausal  1.71    0.157     3.41    6.47e- 4  1.26    2.32
## 7 tumor_stage        1.09    0.0841    1.01    3.11e- 1  0.923   1.28
## 8 three_gene_classifier_subtype 0.880   0.214     -0.596   5.51e- 1  0.579   1.34
## 9 three_gene_classifier_subtype 0.752   0.219     -1.30    1.94e- 1  0.490   1.16
## 10 three_gene_classifier_subtype 0.664   0.292     -1.40    1.61e- 1  0.375   1.18
## # i 29 more rows

```

The multivariable Cox proportional hazards model evaluates how demographic, tumor, and treatment characteristics jointly affect overall survival among METABRIC breast cancer patients. After adjustment for all covariates, traditional prognostic factors such as older age at diagnosis, larger tumor size, a higher number of positive lymph nodes, and a higher Nottingham prognostic index generally show associations compatible with increased hazard of death, indicating worse survival as these measures of disease burden increase.

Tumor-related categorical variables, including tumor stage, three_gene_classifier_subtype (ER/HER2 profile), pam50_claudin_low_subtype, tumor_other_histologic_subtype, cancer_type_detailed, and integrative_cluster, capture biologic and histologic differences between tumors and contribute additional information on risk. Treatment variables such as radio_therapy, along with menopausal status and type_of_breast_surgery, help describe differences in management patterns and their impact on survival. The inclusion of a radio_therapy × tumor_size interaction allows the association between tumor size and mortality risk to differ between patients who did and did not receive radiotherapy.

Hazard ratios from this model represent the multiplicative change in the instantaneous risk of death associated with each covariate, holding all others constant. The directions of effect are broadly consistent with clinical expectations; however, these interpretations rely on the proportional hazards assumption, which is examined in the next section.

Proportional hazards (PH) diagnostics

```
# Residual diagnostics
par(mfrow = c(2, 2))

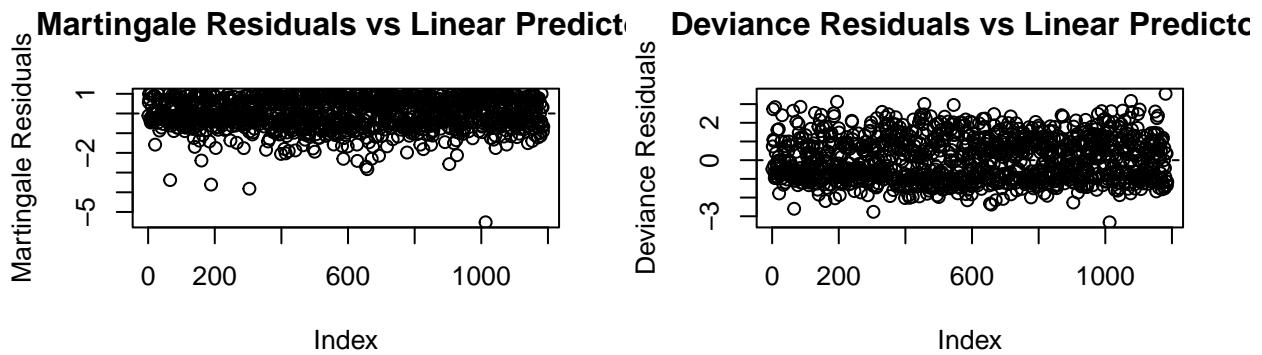
plot(
  resid(final_inter_mod, type = "martingale"),
  main = "Martingale Residuals vs Linear Predictor",
  ylab = "Martingale Residuals"
)
abline(h = 0, lty = 2)

plot(
  resid(final_inter_mod, type = "deviance"),
  main = "Deviance Residuals vs Linear Predictor",
  ylab = "Deviance Residuals"
)
abline(h = 0, lty = 2)

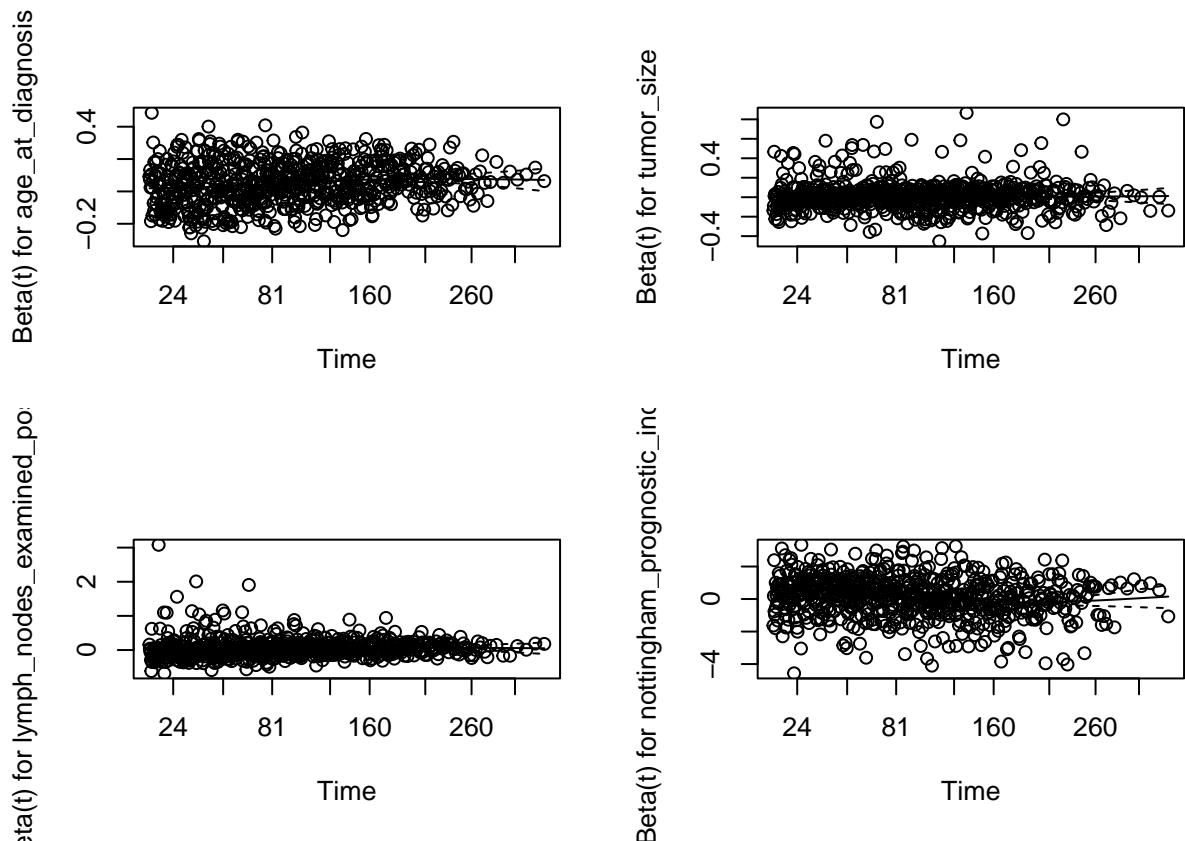
# Schoenfeld residual PH test
ph_test <- cox.zph(final_inter_mod)
ph_test
```

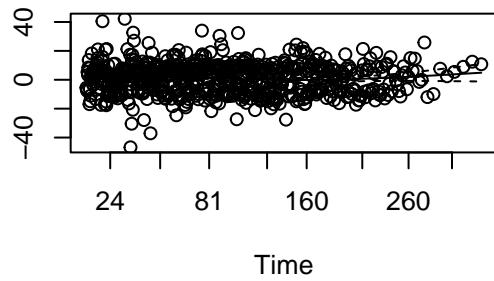
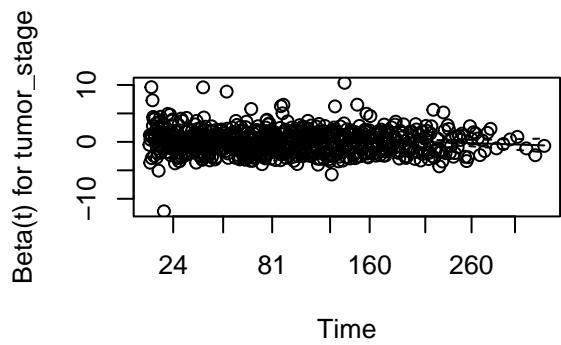
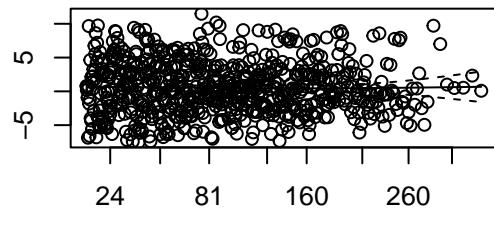
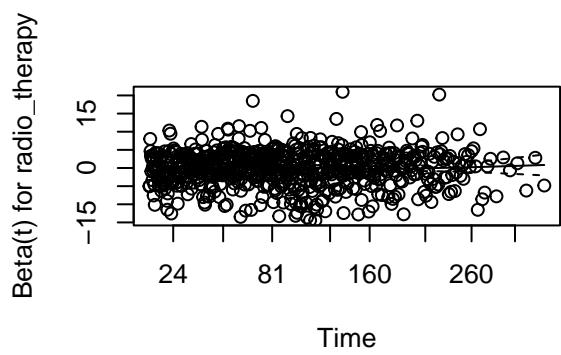
```
##                                chisq df      p
## age_at_diagnosis            56.2469  1 6.4e-14
## tumor_size                  2.6581  1 0.10303
## lymph_nodes_examined_positive 4.1983  1 0.04046
## nottingham_prognostic_index 36.1266  1 1.8e-09
## radio_therapy                0.8096  1 0.36825
## inferred_menopausal_state   26.2766  1 3.0e-07
## tumor_stage                 10.9739  1 0.00092
## three_gene_classifier_subtype 48.0542  3 2.1e-10
## type_of_breast_surgery       0.0169  1 0.89668
## cancer_type_detailed        8.7272  4 0.06829
## pam50_claudin_low_subtype    54.7164  6 5.3e-10
## tumor_other_histologic_subtype 1.0153  2 0.60190
## integrative_cluster          73.8193 10 8.1e-12
## mutation_count                0.1202  1 0.72884
## tumor_size:radio_therapy     1.2559  1 0.26242
## GLOBAL                         139.5229 35 2.0e-14
```

```
# Schoenfeld residual plots
par(mfrow = c(2, 2))
```

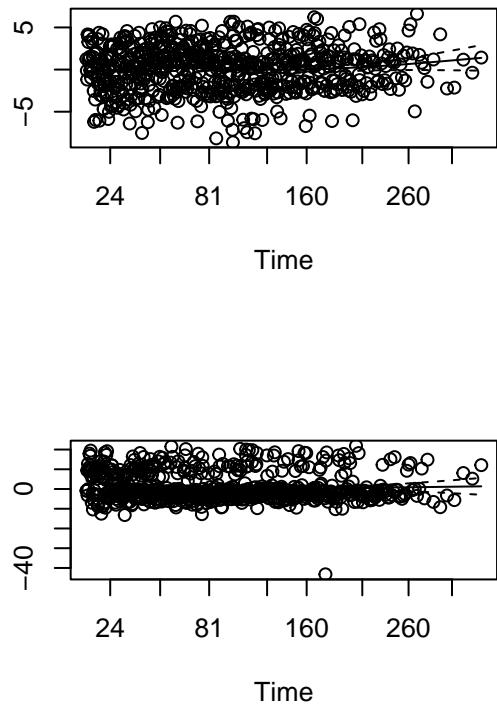


```
plot(ph_test)
```

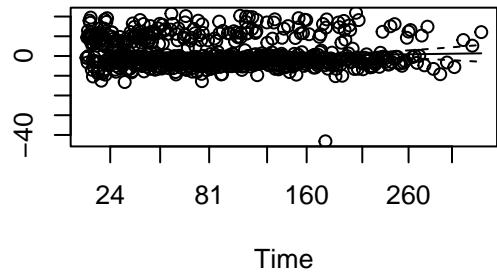




Beta(t) for type_of_breast_surgery

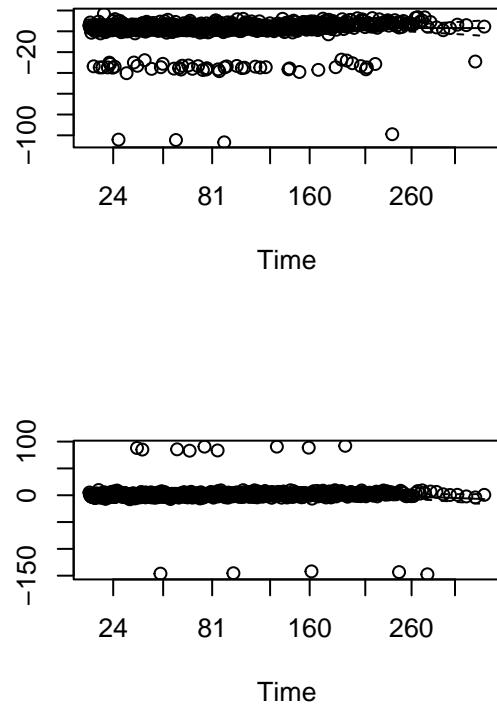


Beta(t) for pam50_claudin_low_subtype

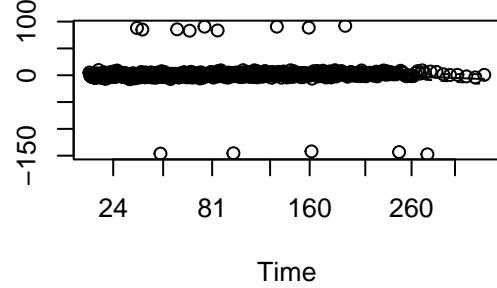


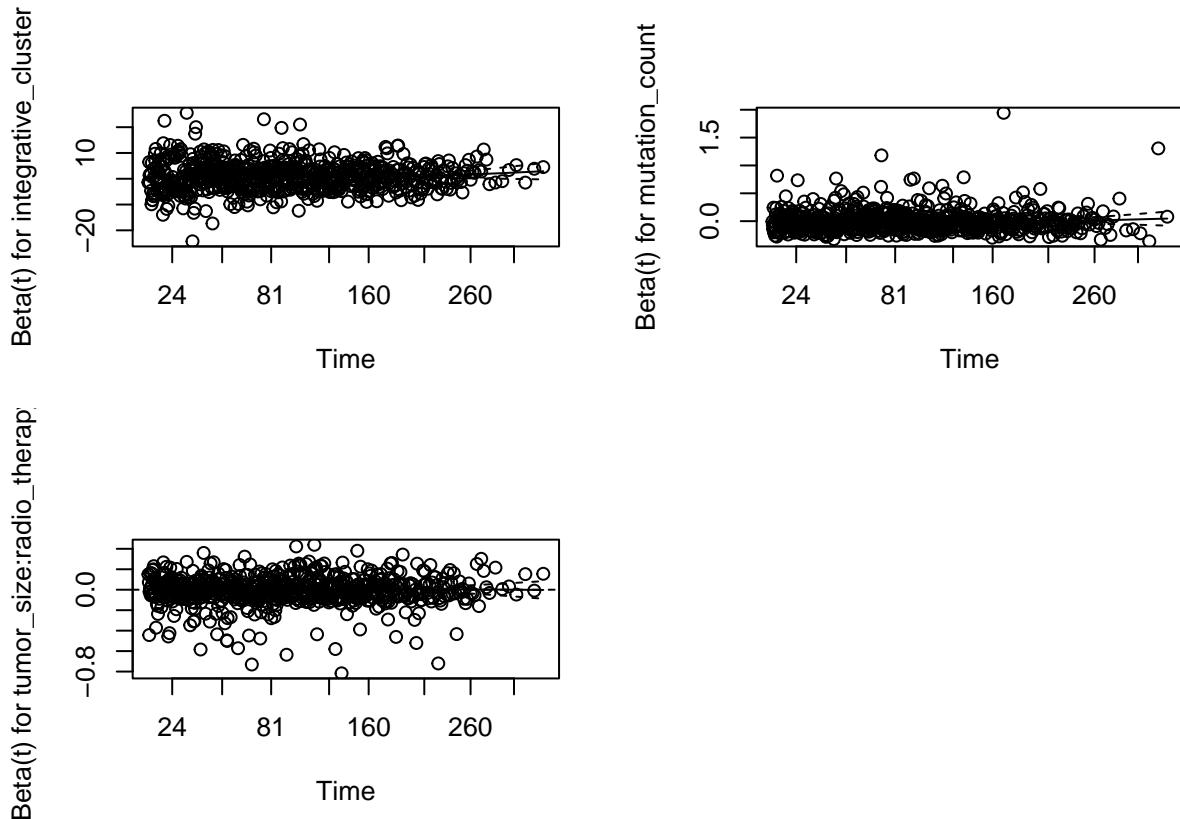
```
abline(h = 0, lty = 2)
par(mfrow = c(1, 1))
```

Beta(t) for cancer_type_detailed



Beta(t) for tumor_other_histologic_subtype





Martingale and deviance residual plots do not reveal severe global model misfit, although a small number of influential observations and mild departures from perfect linearity are apparent. These diagnostics mainly suggest that the overall functional form of the model is reasonable but not flawless.

The Schoenfeld residual test provides a more direct assessment of the proportional hazards assumption. The global test, along with several covariate-specific tests, indicates statistically significant time-varying effects for some predictors. In particular, variables such as age_at_diagnosis, radiotherapy, tumor characteristics, and their interaction ($\text{radio_therapy} \times \text{tumor_size}$) show patterns in their Schoenfeld residual plots that deviate from a flat horizontal line, suggesting that their effects on the hazard of death change over time rather than remaining constant.

Because the proportional hazards assumption is a key requirement of the Cox model, these findings imply that hazard ratios should be interpreted with some caution. To investigate robustness under an alternative modeling framework that does not require proportional hazards, we fit accelerated failure time (AFT) models as a sensitivity analysis.

AFT sensitivity analysis

```
aft_weibull <- survreg(
  Surv(
    overall_survival_months,
    overall_survival_status == "Deceased"
  ) ~
  age_at_diagnosis +
  tumor_size +
  lymph_nodes_examined_positive +
```

```

nottingham_prognostic_index +
radio_therapy +
inferred_menopausal_state +
tumor_stage +
three_gene_classifier_subtype +
type_of_breast_surgery +
cancer_type_detailed +
pam50_claudin_low_subtype +
tumor_other_histologic_subtype +
integrative_cluster +
mutation_count +
log(tumor_size):radio_therapy +
age_at_diagnosis:radio_therapy,
data = filtered_data,
dist = "weibull"
)

# Compare alternative distributions
aft_lognormal <- update(aft_weibull, dist = "lognormal")
aft_loglogistic <- update(aft_weibull, dist = "loglogistic")

AIC(aft_lognormal)

## [1] 8307.446

AIC(aft_weibull)

## [1] 8260.395

AIC(aft_loglogistic)

## [1] 8266.172

# Extract time ratios with confidence intervals for Weibull model
aft_ci <- exp(confint(aft_weibull))
aft_tr <- exp(coef(aft_weibull))

aft_table <- data.frame(
  Variable = names(aft_tr),
  Time_Ratio = aft_tr,
  CI_Lower = aft_ci[, 1],
  CI_Upper = aft_ci[, 2]
)

aft_table

##
## (Intercept)
## age_at_diagnosis
## tumor_size
## lymph_nodes_examined_positive
lymph_

```

```

## nottingham_prognostic_index
## radio_therapyYes
## inferred_menopausal_statePre
## tumor_stage
## three_gene_classifier_subtypeER+/HER2- High Prolif
## three_gene_classifier_subtypeER+/HER2- Low Prolif
## three_gene_classifier_subtypeHER2+
## type_of_breast_surgeryMastectomy
## cancer_type_detailedBreast Invasive Ductal Carcinoma
## cancer_type_detailedBreast Invasive Lobular Carcinoma
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma
## pam50_claudin_low_subtypeclaudin-low
## pam50_claudin_low_subtypeHer2
## pam50_claudin_low_subtypeLumA
## pam50_claudin_low_subtypeLumB
## pam50_claudin_low_subtypeNC
## pam50_claudin_low_subtypeNormal
## tumor_other_histologic_subtypeLobular
## tumor_other_histologic_subtypeMedullary
## tumor_other_histologic_subtypeMixed
## tumor_other_histologic_subtypeMucinous
## tumor_other_histologic_subtypeOther
## tumor_other_histologic_subtypeTubular/ cribriform
## integrative_cluster10
## integrative_cluster2
## integrative_cluster3
## integrative_cluster4ER-
## integrative_cluster4ER+
## integrative_cluster5
## integrative_cluster6
## integrative_cluster7
## integrative_cluster8
## integrative_cluster9
## mutation_count
## radio_therapyNo:log(tumor_size)
## radio_therapyYes:log(tumor_size)
## age_at_diagnosis:radio_therapyYes
##
## (Intercept)
## age_at_diagnosis
## tumor_size
## lymph_nodes_examined_positive
## nottingham_prognostic_index
## radio_therapyYes
## inferred_menopausal_statePre
## tumor_stage
## three_gene_classifier_subtypeER+/HER2- High Prolif
## three_gene_classifier_subtypeER+/HER2- Low Prolif
## three_gene_classifier_subtypeHER2+
## type_of_breast_surgeryMastectomy
## cancer_type_detailedBreast Invasive Ductal Carcinoma
## cancer_type_detailedBreast Invasive Lobular Carcinoma
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma
## pam50_claudin_low_subtypeclaudin-low
## pam50_claudin_low_subtypeHer2
## pam50_claudin_low_subtypeLumA
## pam50_claudin_low_subtypeLumB
## pam50_claudin_low_subtypeNC
## pam50_claudin_low_subtypeNormal
## tumor_other_histologic_subtypeLobular
## tumor_other_histologic_subtypeMedullary
## tumor_other_histologic_subtypeMixed
## tumor_other_histologic_subtypeMucinous
## tumor_other_histologic_subtypeOther
## tumor_other_histologic_subtypeTubular/ cribriform
## integrative_cluster10
## integrative_cluster2
## integrative_cluster3
## integrative_cluster4ER-
## integrative_cluster4ER+
## integrative_cluster5
## integrative_cluster6
## integrative_cluster7
## integrative_cluster8
## integrative_cluster9
## mutation_count
## radio_therapyNo:log(tumor_size)
## radio_therapyYes:log(tumor_size)
## age_at_diagnosis:radio_therapyYes
## Time_Ratio
3.981209e+08
9.660603e-01
1.001310e+00
9.621810e-01
9.162981e-01
1.180931e+00
6.926896e-01
9.730263e-01
1.100052e+00
1.240856e+00
1.356524e+00
9.562415e-01
1.336101e-05
1.551825e-05
2.017107e-05

```

```

## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma 1.300091e-05
## pam50_claudin_low_subtypeclaudin-low 1.227576e+00
## pam50_claudin_low_subtypeHer2 9.263517e-01
## pam50_claudin_low_subtypeLumA 1.155411e+00
## pam50_claudin_low_subtypeLumB 1.142379e+00
## pam50_claudin_low_subtypeNC 1.957232e+00
## pam50_claudin_low_subtypeNormal 8.807838e-01
## tumor_other_histologic_subtypeLobular NA
## tumor_other_histologic_subtypeMedullary 7.454257e-01
## tumor_other_histologic_subtypeMixed NA
## tumor_other_histologic_subtypeMucinous NA
## tumor_other_histologic_subtypeOther NA
## tumor_other_histologic_subtypeTubular/ cribriform 1.382164e+00
## integrative_cluster10 1.293235e+00
## integrative_cluster2 8.970724e-01
## integrative_cluster3 9.756006e-01
## integrative_cluster4ER- 1.080654e+00
## integrative_cluster4ER+ 1.068352e+00
## integrative_cluster5 6.520211e-01
## integrative_cluster6 1.027033e+00
## integrative_cluster7 1.032745e+00
## integrative_cluster8 1.005275e+00
## integrative_cluster9 1.013440e+00
## mutation_count 9.955238e-01
## radio_therapyNo:log(tumor_size) 7.446536e-01
## radio_therapyYes:log(tumor_size) 7.266791e-01
## age_at_diagnosis:radio_therapyYes 1.001072e+00
##
## CI_Lower
## (Intercept) 0.0000000
## age_at_diagnosis 0.9577066
## tumor_size 0.9942822
## lymph_nodes_examined_positive 0.9468032
## nottingham_prognostic_index 0.8527211
## radio_therapyYes 0.4282061
## inferred_menopausal_statePre 0.5583204
## tumor_stage 0.8627505
## three_gene_classifier_subtypeER+/HER2- High Prolif 0.8189749
## three_gene_classifier_subtypeER+/HER2- Low Prolif 0.9170685
## three_gene_classifier_subtypeHER2+ 0.9048474
## type_of_breast_surgeryMastectomy 0.8252283
## cancer_type_detailedBreast Invasive Ductal Carcinoma 0.0000000
## cancer_type_detailedBreast Invasive Lobular Carcinoma 0.0000000
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma 0.0000000
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma 0.0000000
## pam50_claudin_low_subtypeclaudin-low 0.9180968
## pam50_claudin_low_subtypeHer2 0.6831930
## pam50_claudin_low_subtypeLumA 0.8383571
## pam50_claudin_low_subtypeLumB 0.8265686
## pam50_claudin_low_subtypeNC 0.4693102
## pam50_claudin_low_subtypeNormal 0.6147883
## tumor_other_histologic_subtypeLobular NA
## tumor_other_histologic_subtypeMedullary 0.4603267
## tumor_other_histologic_subtypeMixed NA
## tumor_other_histologic_subtypeMucinous NA

```

## tumor_other_histologic_subtypeOther	NA
## tumor_other_histologic_subtypeTubular/ cribriform	0.7337697
## integrative_cluster10	0.9230238
## integrative_cluster2	0.6442743
## integrative_cluster3	0.7351044
## integrative_cluster4ER-	0.7249431
## integrative_cluster4ER+	0.8004487
## integrative_cluster5	0.4358867
## integrative_cluster6	0.7446676
## integrative_cluster7	0.7719673
## integrative_cluster8	0.7638130
## integrative_cluster9	0.7540969
## mutation_count	0.9833782
## radio_therapyNo:log(tumor_size)	0.5702929
## radio_therapyYes:log(tumor_size)	0.5350255
## age_at_diagnosis:radio_therapyYes	0.9917696
##	CI_Upper
## (Intercept)	Inf
## age_at_diagnosis	0.9744869
## tumor_size	1.0083881
## lymph_nodes_examined_positive	0.9778086
## nottingham_prognostic_index	0.9846152
## radio_therapyYes	3.2568400
## inferred_menopausal_statePre	0.8593971
## tumor_stage	1.0973974
## three_gene_classifier_subtypeER+/HER2- High Prolif	1.4775971
## three_gene_classifier_subtypeER+/HER2- Low Prolif	1.6789631
## three_gene_classifier_subtypeHER2+	2.0336673
## type_of_breast_surgeryMastectomy	1.1080543
## cancer_type_detailedBreast Invasive Ductal Carcinoma	Inf
## cancer_type_detailedBreast Invasive Lobular Carcinoma	Inf
## cancer_type_detailedBreast Invasive Mixed Mucinous Carcinoma	Inf
## cancer_type_detailedBreast Mixed Ductal and Lobular Carcinoma	Inf
## pam50_claudin_low_subtypeclaudin-low	1.6413768
## pam50_claudin_low_subtypeHer2	1.2560543
## pam50_claudin_low_subtypeLumA	1.5923699
## pam50_claudin_low_subtypeLumB	1.5788519
## pam50_claudin_low_subtypeNC	8.1625280
## pam50_claudin_low_subtypeNormal	1.2618655
## tumor_other_histologic_subtypeLobular	NA
## tumor_other_histologic_subtypeMedullary	1.2070981
## tumor_other_histologic_subtypeMixed	NA
## tumor_other_histologic_subtypeMucinous	NA
## tumor_other_histologic_subtypeOther	NA
## tumor_other_histologic_subtypeTubular/ cribriform	2.6035126
## integrative_cluster10	1.8119335
## integrative_cluster2	1.2490627
## integrative_cluster3	1.2947774
## integrative_cluster4ER-	1.6109025
## integrative_cluster4ER+	1.4259206
## integrative_cluster5	0.9753258
## integrative_cluster6	1.4164671
## integrative_cluster7	1.3816161
## integrative_cluster8	1.3230698

## integrative_cluster9	1.3619740
## mutation_count	1.0078194
## radio_therapyNo:log(tumor_size)	0.9723231
## radio_therapyYes:log(tumor_size)	0.9869858
## age_at_diagnosis:radio_therapyYes	1.0104622

Given the evidence of proportional hazards violations in the Cox model, we fitted accelerated failure time (AFT) models as a robustness check. AFT models describe how covariates multiply survival time directly, rather than affecting the hazard, and therefore do not require proportional hazards.

Using the same core predictors as in the Cox model, along with interactions involving log-transformed tumor_size and radiotherapy and age_at_diagnosis with radiotherapy, we fitted Weibull, lognormal, and log-logistic AFT models. Comparison of AIC values identified the Weibull distribution as the best-fitting parametric form among the three candidate models.

In the AFT framework, exponentiated coefficients are interpreted as time ratios: a time_ratio greater than 1 indicates longer expected survival, whereas a time_ratio less than 1 indicates shorter expected survival, holding other factors constant. Consistent with the Cox model, factors reflecting greater tumor burden or worse prognosis (e.g., larger tumor_size, more lymph_nodes_examined_positive, higher nottingham_prognostic_index, and more advanced tumor_stage) are associated with shorter survival times. The general agreement between the Cox and Weibull AFT results suggests that the main substantive conclusions about key prognostic factors are robust, even when relaxing the proportional hazards assumption.

Table 1: **Table 1. Baseline characteristics by overall survival status**

Characteristic	N	Deceased N = 1,144 ¹	Living N = 834 ¹	p-
age_at_diagnosis	1,978			<
Mean (SD)		64 (13)	57 (11)	<
tumor_size	1,955			<
Mean (SD)		28 (16)	23 (13)	<
Unknown		16	7	<
lymph_nodes_examined_positive	1,904			<
Mean (SD)		2.6 (4.8)	1.2 (2.7)	<
Unknown		40	34	<
mutation_count	1,859			<
Mean (SD)		6.0 (4.5)	5.3 (3.3)	<
Unknown		55	64	<
nottingham_prognostic_index	1,977			<
Mean (SD)		4.15 (1.20)	3.83 (1.08)	<
Unknown		1	0	<
type_of_breast_surgery	1,953			<
Breast Conserving		368 (33%)	416 (50%)	<
Mastectomy		759 (67%)	410 (50%)	<
Unknown		17	8	<
cancer_type_detailed	1,978			<
Breast		5 (0.4%)	12 (1.4%)	<
Breast Invasive Ductal Carcinoma		891 (78%)	646 (77%)	<
Breast Invasive Lobular Carcinoma		86 (7.5%)	60 (7.2%)	<
Breast Invasive Mixed Mucinous Carcinoma		10 (0.9%)	13 (1.6%)	<
Breast Mixed Ductal and Lobular Carcinoma		132 (12%)	79 (9.5%)	<
Invasive Breast Carcinoma		18 (1.6%)	24 (2.9%)	<
Metaplastic Breast Cancer		2 (0.2%)	0 (0%)	<
pam50_claudin_low_subtype	1,977			<
Basal		115 (10%)	94 (11%)	<
claudin-low		94 (8.2%)	121 (15%)	<
Her2		157 (14%)	67 (8.0%)	<
LumA		378 (33%)	322 (39%)	<
LumB		315 (28%)	160 (19%)	<
NC		5 (0.4%)	1 (0.1%)	<
Normal		79 (6.9%)	69 (8.3%)	<
Unknown		1	0	<
neoplasm_histologic_grade	1,893			<
1		76 (7.0%)	93 (12%)	<
2		432 (40%)	339 (42%)	<
3		579 (53%)	374 (46%)	<
Unknown		57	28	<
tumor_other_histologic_subtype	1,936			<
Ductal/NST		871 (77%)	620 (77%)	<
Lobular		86 (7.6%)	60 (7.4%)	<
Medullary		14 (1.2%)	11 (1.4%)	<
Metaplastic		2 (0.2%)	0 (0%)	<
Mixed	43	132 (12%)	79 (9.8%)	<
Mucinous		10 (0.9%)	13 (1.6%)	<
Other		5 (0.4%)	12 (1.5%)	<
Tubular/cribiform		6 (0.5%)	15 (1.8%)	<