# Biostatistical Informatics Assignment 2

Steven Mitchell / 40175882

## 1. Summarise the composition of the data

In this study, 686 patients with a median age of 53 years (range: 30-80) were included. Of these patients, 41% were reported as pre-menopausal whilst a majority (404) where post-menopausal. Hormone therapy was administered to just over 1 third of the cohort (252), while 432 did not receive this treatment. The median tumour size reported was 25 mm with a large range spanning 82mm (8-90 mm). Nearly 2 of every 3 patients reported a tumour grade of 2 (64%) with the remainder mostly recorded as grade 3 (24%) or 1 (12%). The median survival time for this group of patients was 2400 days (range: 100-2668 days) with 189 deaths (events) recorded.

TABLE 1

Data composition summary

| Category | Level | N | n |
|---|---|---|---|
| Age (years) | *Median: 53 (Range: 30 - 80)* | 686 | 189 |
| Menopause status | *Pre* | 282 | 58 |
| | *Post* | 404 | 131 |
| Hormone therapy | *Yes* | 254 | 71 |
| | *No* | 432 | 118 |
| Tumour size | *Median: 25 (Range: 8 - 90)* | 686 | 189 |
| Tumour Grade | *1* | 84 | 12 |
| | *2* | 441 | 118 |
| | *3* | 161 | 59 |
| Survival (Days) | *Median: 2400 (Range: 100 - 2668)* | 686 | 189 |

N = sub-total
n = number of events

## 2. Considering all patients develop a FULL multivariate model from the clinico-pathological variables provided. Select the variables which best explain survival to establish a FINAL multivariate model.

**Methods**

Cox proportional hazards regression analysis was used to determine the relationship between survival and the various factors recorded in the study such as age, menopause status, hormone therapy, tumour size, and tumour grade. Stepwise selection was then performed to identify the best-fit model. Additionally, the Akaike information criterion (AIC) was calculated for both the full Cox model and the stepwise selection processed model to compare the fit, with a lower value indicating a better fit. Visual models were also developed to glean further understanding in which variables best explain survival within the multivariate model. Kaplan-Meier curves were used to demonstrate the relationship between survival and predictor variables, which was accompanied by a log-rank test to compare survival between groups. The p-values from the log-rank tests are displayed to indicate significance where p <= 0.05 is considered significant.

**Results**

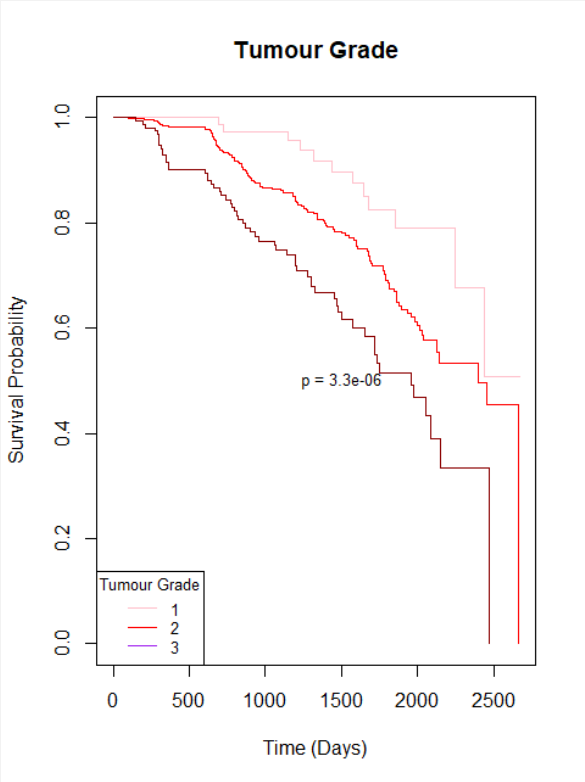
Table 2 shows both the full model including all predictors and the reduced model following stepwise selection to show only the significant predictors of menopause status, tumour size, and tumour grade. The first (full model) shows the hazard ratio associated with a one-year increase in age is 1.003, but this effect is not statistically significant (p = 0.791). Conversely, patients who have reached menopause have a hazard ratio of 1.61 compared to those who have not reached menopause (p = 0.049), suggesting that menopause status is a significant predictor of survival. Hormone therapy, unlike the other variables shows a negative hazard ratio of 0.834, but this effect is not statistically significant (p = 0.238). Increase in tumour size is associated with a hazard ratio of 1.015 and this effect is statistically significant (p = 0.003), along with high tumour grade which corresponds to a very high jump in hazard ratio of 1.883 (p < 0.001).

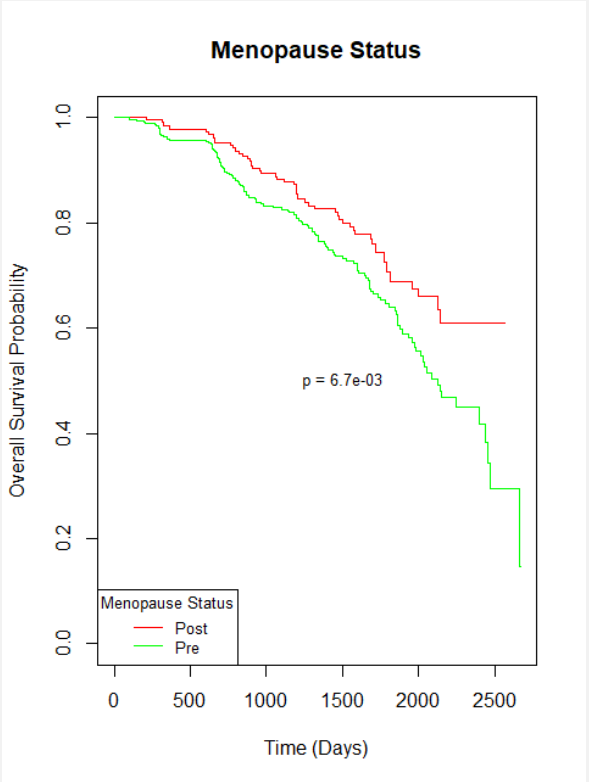
TABLE 2

Cox proportional hazards regression models for potential predictors of survival, showing both the full model and adjustment following stepwise selection.

| Variable | Hazard ratio | P-value |
|---|---|---|
| *Full model* | | |
| Age | 1.003 | 0.791 |
| Menopause status | 1.610 | 0.049 |
| Hormone therapy | 0.834 | 0.238 |
| Tumour size | 1.015 | 0.003 |
| Tumour grade | 1.883 | *0.000* |
| *Final model(stepwise)* | | |
| Menopause Status | 1.637 | 0.002 |
| Tumour Size | 1.016 | 0.002 |
| Tumour Grade | 1.900 | *0.000* |

When stepwise selection was used to focus on the most important predictors of survival, it is again demonstrated the significance of menopause status, tumour size and tumour grade. In this case, the C-index is 0.653, which indicates that the model has a fair level of predictive accuracy. Furthermore, the likelihood ratio test, Wald test and Score (logrank) test all provided extremely very small p-values less than 0.0001, indicating that the model as a whole is statistically significant and can be used to predict survivability based on the three predictors. The AIC value for the reduced model was 2137.864 versus 2140.404 for the full model, indicating the final model (stepwise) provides a better balance between fit and parsimony, and may be more appropriate to use for prediction. The significance of these identified variables in terms of survivability predictions is further visualised via the KM curves shown in *figure 1*.



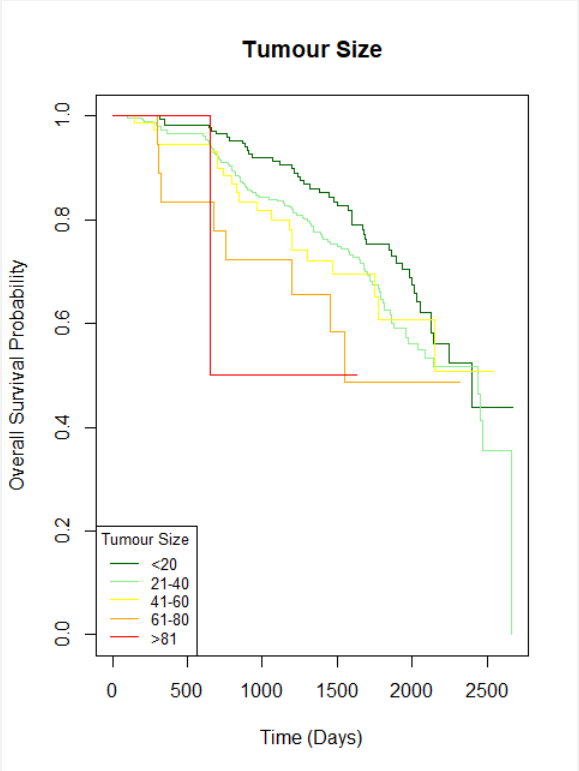




Fig. 1

Kaplan-Meier curves showing the overall survival probability against tumour grade, menopause status and tumour size, each including a p-value to show the significance of the relationship between group differences and survivability.

## 3. A collaborator is interested if age behaves differently in the pre-menopausal and post-menopausal groups of patients. Compare and contrast the behaviour of age (using univariate and multivariate) in both groups.

**Methods**

Data was split into the pre-menopausal and post-menopausal groups where the behaviour of age could be compared using both univariate and multivariate methods. Univariate methods examined the relationship between a single predictor variable and the outcome variable, while multivariate methods examined the relationship between multiple predictor variables and the outcome variable while controlling for other variables. Boxplots were created to show the age distribution of each group and compare the median, quartiles, and outliers. To answer the collaborator's question more directly, the analysis focused on the relationship between age and survival in pre- and post-menopausal patients. Specifically, it was tested whether older pre-menopausal (>group mean) patients had lower survival rates than younger post-menopausal patients (<group mean) using a Welch Two Sample t-test (normality of data tested using Shapiro-Wilk test for normality). Further univariate analysis was conducted for each group to investigate the relationship between age and survival with a Cox proportional hazards model for age and survival separately for each group. Finally, the Wilcoxon rank-sum test was conducted to compare age between patients who survived and those who did not survive separately for each group. To further visualise the relationship, Kaplan-Meier curves were plotted to show the probability of survival over time in different age groups. A log-rank test was also performed to determine whether there are significant differences in survival between age groups in both pre-menopausal and post-menopausal groups.

A multivariate Cox proportional hazards model was generated to include age and menopausal status as predictors of survival. Next, an interaction test of age and menopausal status on survival was performed to assess whether the effect of one predictor variable on the outcome variable differs depending on the level of another predictor variable. A linear regression model with interaction term was fitted to the data to assess whether the effect of age on the outcome variable differs between the pre-menopausal and post-menopausal groups.

**Results**

*Figure 2* illustrates the age distribution between the pre- and post-menopausal groups, which as expected shows the pre-menopausal group having a significantly lower average age (44) versus the post-menopausal group (59 years). When answering an initial research question of whether older pre-menopausal (> group mean) patients have lower or higher survival than younger post-menopausal (<group mean) patients, it was determined via a Welch Two Sample t-test that sample estimates for the mean survival of both groups (1272.396 and 1327.321) suggested the postmenopausal group had a higher survival rate (data passed normality assumption p = < 0.05). However, the statistical analysis does not provide enough evidence to reject the null hypothesis (p = 0.2).

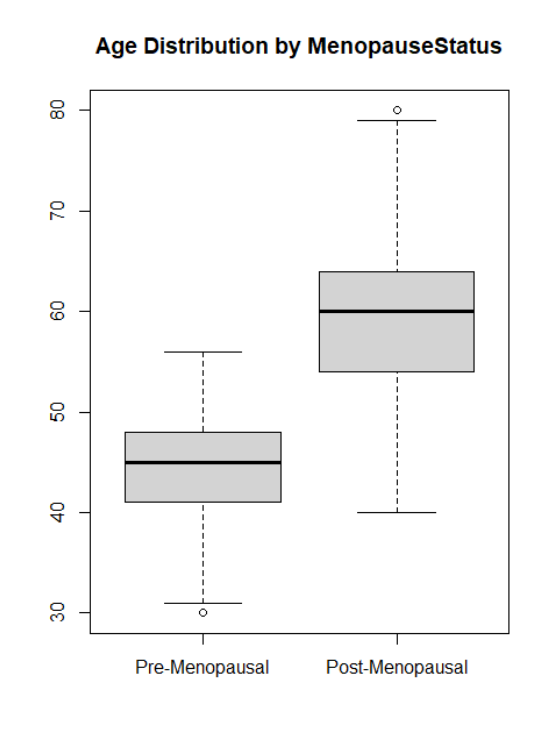



Fig. 2

Boxplot comparing age distribution in the pre-menopause and post-menopause cohorts.

Cox proportional hazards regression model in first the pre and then post-menopausal group was utilised to investigate the relationship between age and survival. The main coefficient of interest is for age which in the pre-menopausal group was estimated at 0.02531 indicating that increasing age is associated with a slightly higher hazard of experiencing an event. However, the coefficient is not statistically significant with a p-value of 0.297. In the post-menopausal group the age coefficient ( -0.007157) indicated that an increase in age had a lower hazard of event. However, again the conclusion was found to be not statistically significant with a p-value of 0.587, a concordance score indicating poor predictive power and likelihood ratio, Wald, and Score tests all having p-values greater than 0.05. Therefore, initial univariate testing indicated age was not a significant predictor of survival in the model. In addition to the Cox model, two Wilcoxon rank-sum tests were conducted to compare the age variable between two groups, again producing non-significant p-values of 0.3815 and 0.7429, respectively. *Figure 3* also highlights these findings, with accompanying log-rank outputs for each group of p = 0.6 and p = 1.

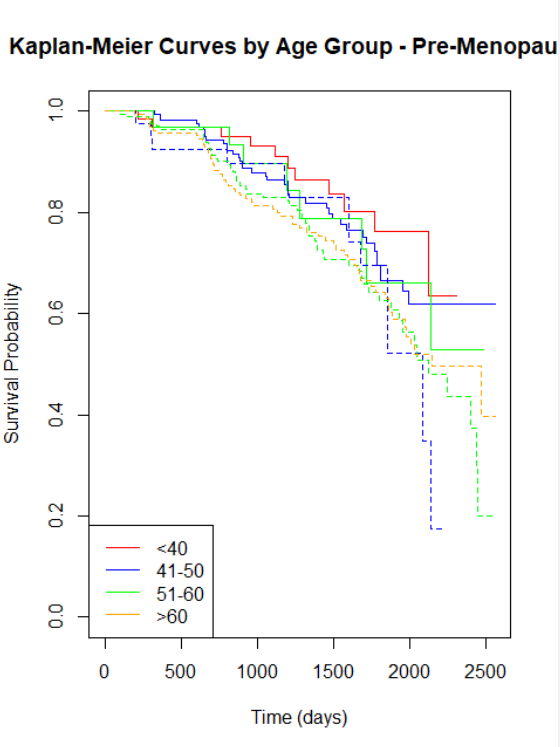



Fig. 3

KM curves for age groups and associated survival.

Multivariate testing using a Cox proportional hazards model with predictor covariates as age and menopause status. The output suggests that menopause status is marginally significant in predicting survival (p=0.0779) with women who are post-menopausal having a 1.52 times higher hazard of event, but not to an extent considered significant (p<0.05). The concordance index of 0.558 suggests that the model has moderate predictive accuracy, and when coupled with the findings of likelihood ratio, Wald, and score tests that indicate that the full model is significant with p-values less than 0.05, it shows more investigation is required to fully understand the dataset. An interaction term between age and menopause status was added to the model which returned similar results wherein age, menopause status and the interaction term were all non-significant (p = 0.307, 0.1780, 0.295 respectively). Linear regression models for both event and survival were generated using the same three predictors, however, all three were found not to be statistically significant based on their respective p-values (>0.05). Additionally, the adjusted R-squared value for survival was negative, indicating that the model could not explain the variability in the data. Overall, the generated models could not provide significant evidence to establish a relationship between age and menopause status as predictors for survivability.

## 4. Your collaborator defines "good" survival as those patients who have survived beyond five years and those with "poor" survival as dying before the first year.

### (a) Using expression levels of genes 1 to 5 in a second data file (assignOct4.txt), which genes, if any, have different expression levels between the "good survival" and "poor survival" patient groups?

**Methods**

To identify any significant difference in gene expression levels good and poor survival patient groups the Wilcoxon rank sum test was used ($p < 0.05$ considered significant). Box plots were also created to visualise expression levels for each survival group. An additional Bonferroni correction was added to account for the increased chance of obtaining false positives due to multiple testing (corrected significant p-value = 0.01). Because the dataset includes multiple variables an analysis of variance (ANOVA) was also performed with a post-hoc Tukey's Honest Significant Difference (HSD) test to report and identify significantly different means.

TABLE 3

Wilcoxon rank sum and ANOVA results for survival vs gene expression. Significant values are shown in bold.

| Gene | Wilcoxon | ANOVA |
|---|---|---|
| 1 | 0.048 | 0.102 |
| 2 | 0.042 | 0.068 |
| 3 | 0.923 | 0.923 |
| 4 | 0.611 | 0.822 |
| 5 | 0.019 | **0.015** |

**Results**

Before Bonferroni correction, it was observed genes 1, 2 and 5 had significantly different expression profiles within the good and poor survival cohorts. However, following correction, none of genes had significantly different expression levels, with gene 5 the closest to the corrected p-value limit, as shown in *Table 3*. ANOVA output indicated Genes 1-4 do not show a significant difference in expression levels between survival groups, however, Gene 5 showed significantly lower expression in

the poor survival group (adjusted p = 0.015). Results were further visualised using box plots, shown in *figure 4*.

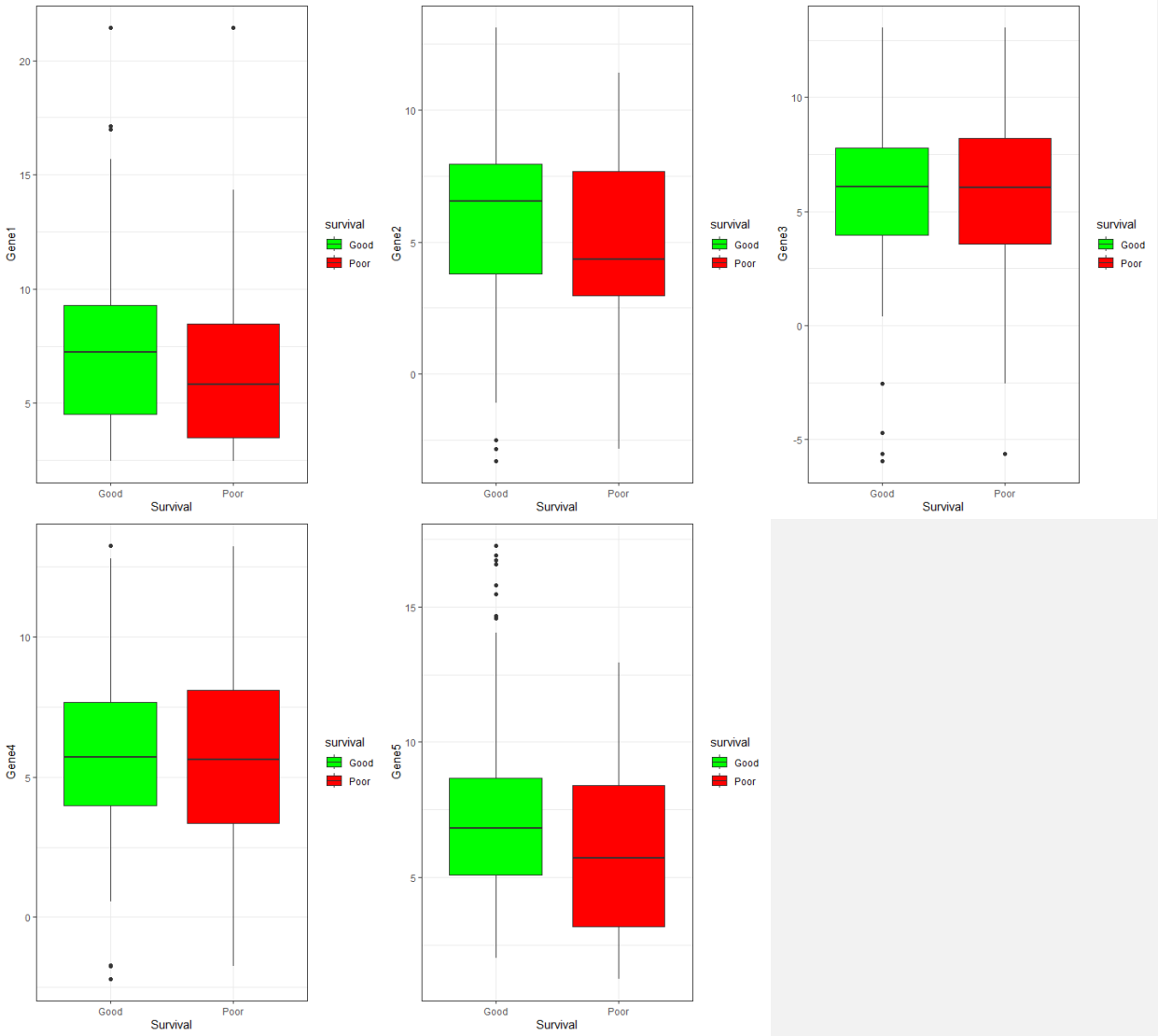



Fig. 4

Boxplots for Gene expression vs survival group. It is visually clear there is little variance in genes 3 and 4, whereas 1, 2 and 5 require further testing to understand if differences are significant.

## (b) Use a semi-parametric (univariate) method to consider the relationship of each gene to overall survival. Comment on the similarities/differences with the results in part a).

**Methods**

A Cox proportional hazards regression model was fitted for each gene to determine the relationship of each gene with overall survival. Boxplots were generated to again visualise the distribution of expression levels for each gene, separated by survival status (alive or dead). Similarities and differences between the results for each gene were then evaluated and discussed in comparison to the findings from the previous analysis.

**Results**

Hazard ratios, p-values, and concordance statistics were calculated for each gene as shown in *Table 4*, extracted from each gene's Cox proportional hazards regression model. Comparing the results from *Table 3* supported by the plots in *figures 4 and 5*, some similarities are present. Genes 2, 3 and 4 did not show a significant association with survival in either analysis. In contrast, Gene 1 showed a significant difference (p=0.00795) in expression levels and survival in the Cox model, with a hazard ratio <1 indicative of a good prognostic factor, and the highest concordance value of the genes analysed implying better predictive power. Gene 1, when corrected in the previous analysis, was not shown to be significantly different suggesting it may be a predictor of overall survival independent of association to a defined survival group. Conversely, Gene 5 was notably significant when associated with survival group, but in this analysis did not appear significantly predictive of overall survival. Further investigation may be needed to determine the underlying reasons for these differences.

TABLE 4

Cox proportional hazards regression model outputs for each gene 1-5 for Hazard Ratio, P-value, and Concordance.

| Gene | Hazard Ratio | P-value | Concordance |
|---|---|---|---|
| 1 | 0.951 | **0.008** | 0.571 |
| 2 | 0.995 | 0.822 | 0.537 |
| 3 | 1.024 | 0.228 | 0.512 |
| 4 | 1.023 | 0.329 | 0.508 |
| 5 | 0.965 | 0.171 | 0.538 |

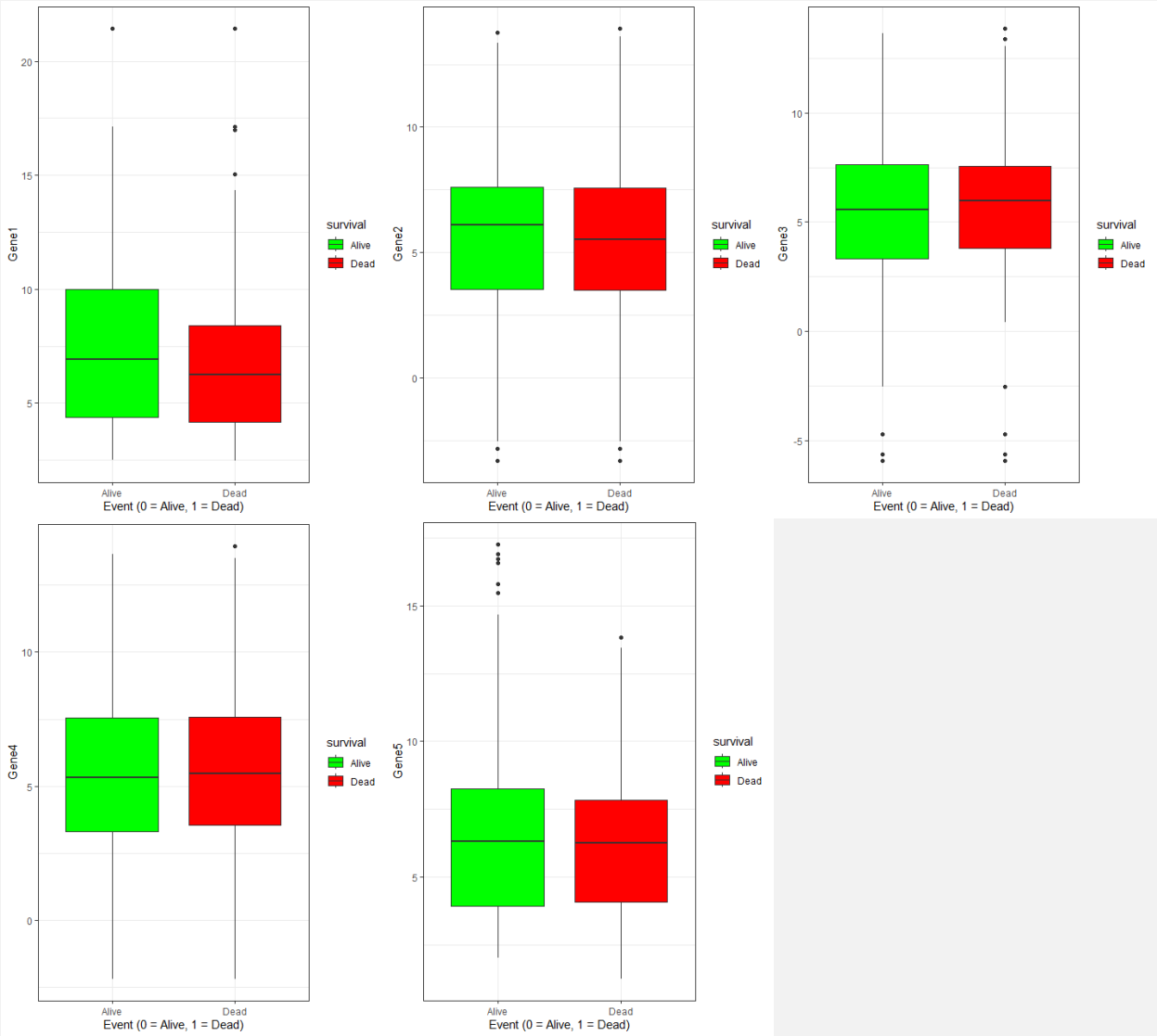



Fig. 5

Boxplots for Gene expression vs event. It is visually less clear cut in comparison to figure 4 data which, if any gene expression has a significant effect on overall survival status, prompting the need for further analysis.

```
# APPENDIX


# READ IN DATASET
  data <-read.table("assignOct3-1.txt", header=TRUE, row.names=1)
  # check read
    head(data)
  # check dimensions
    dim(data)
  # attach
    attach(data)
  # structure check
    str(data)


# QUESTION 1

    # events
      table(data$Event)

    # age
      summary(data$Age)

    # menopause
      table(data$MenopauseStatus)
      table(data$MenopauseStatus, data$Event)

    # hormone therapy
      table(data$HormoneTherapy)
      table(data$HormoneTherapy, data$Event)

    # tumour size
      summary(data$TumourSize)

    # tumour grade
      table(data$TumourGrade)
      table(data$TumourGrade, data$Event)

    # Median survival
      library(survival)
      survfit(Surv(data$Survival, data$Event) ~ 1)
      summary(data$Survival)

# QUESTION 2

    # Cox proportional hazards regression analysis
      model_full <- coxph(Surv(Survival, Event) ~ Age + MenopauseStatus +
                            HormoneTherapy + TumourSize + TumourGrade,
                          data = data)
      model_full
    # stepwise selection
      model_final <- step(model_full)

    # model summaries
      summary(model_final)

    # AIC values check (lower is better fit)
      AIC(model_full)
      AIC(model_final)

    # visuals
        # km tumour grade
          library(survival)
          KM.tumourgrade <- survfit(Surv(Survival, Event) ~ TumourGrade,
                                    data = data)
          diff.test <- survdiff(Surv(Survival, Event) ~ TumourGrade,
                                data = data)
          p.value.TG <- format(diff.test$p, scientific = TRUE, digits = 2)
```

```
  plot(KM.tumourgrade, main = "Tumour Grade", xlab = "Time (Days)",
       ylab = "Survival Probability", col = c("pink", "red", "darkred"),
       lty = 1)
  legend("bottomleft", title = "Tumour Grade", legend = c("1", "2", "3"),
         col = c("pink", "red", "purple"), lty = 1, cex = 0.8)
  text(1500, 0.5, paste("p =", p.value.TG), cex = 0.8)

# km menopause
  KM.menopause <- survfit(Surv(Survival, Event) ~ MenopauseStatus,
                          data = data)
  diff.test <- survdiff(Surv(Survival, Event) ~ MenopauseStatus,
                        data = data)
  p.value.MP <- format(diff.test$p, scientific = TRUE, digits = 2)

  plot(KM.menopause, main = "Menopause Status", xlab = "Time (Days)",
       ylab = "Overall Survival Probability", col = c("red", "green"),
       lty = 1)
  legend("bottomleft", title = "Menopause Status", c("Post", "Pre"),
         lty = 1,
         col = c("red", "green"), cex = 0.8)
  text(1500, 0.5, paste("p =", p.value.MP), cex = 0.8)


# km tumour size
    # break into groups
    data$TumourSizeGroup <- cut(data$TumourSize,
                                c(0, 20, 40, 60, 80, Inf),
                               labels = c("<20", "21-40", "41-60",
                                          "61-80", ">81"))
    # create curves
    my.KMest <- survfit(Surv(Survival, Event) ~ TumourSizeGroup,
                        data = data)
    # plot
    plot(my.KMest, main = "Tumour Size", xlab = "Time (Days)",
       ylab = "Overall Survival Probability",
       col = c("darkgreen", "lightgreen", "yellow", "orange", "red"),
       lty = 1)
    # add legend
    legend("bottomleft", title = "Tumour Size",
           c("<20", "21-40", "41-60", "61-80", ">81"),
         lty = 1,
         col = c("darkgreen", "lightgreen", "yellow", "orange", "red"),
         cex = 0.8)
    # add p-value
    fit <- survdiff(Surv(Survival, Event) ~ TumourSizeGroup, data = data)
    p.value <- format(round(summary(fit)$chisq["pvalue"], 4), nsmall = 4)
    legend("bottomright", paste0("p-value: ", p.value), cex = 0.8,)

# QUESTION 3

  # split data into pre-menopausal and post-menopausal groups
  pre_meno <- subset(data, data$MenopauseStatus == 1)
  post_meno <- subset(data, data$MenopauseStatus == 2)

  # compare behavior of age in pre- and post-menopausal patients
    # measures of central tendency
    mean(pre_meno$Age)
    mean(post_meno$Age)
    # boxplot
    boxplot(pre_meno$Age, post_meno$Age,
            main = "Age Distribution by MenopauseStatus",
            names = c("Pre-Menopausal", "Post-Menopausal"))

    # do older pre menopausal patients have  a lower or higher
    #survival than younger post menopasual patients?
      # filter by age and menopausal status
      oldpremenopausal <- subset(data,
                                 MenopauseStatus == 1 & Age > 45)
```

```
                youngpostmenopausal <- subset(data,
                                              MenopauseStatus == 2 & Age < 59)
              # check normality
                shapiro_test_oldpremenopausal <-
shapiro.test(oldpremenopausal$Survival)
                shapiro_test_youngpostmenopausal <-
shapiro.test(youngpostmenopausal$Survival)
              # print Shapiro-Wilk
                cat("Shapiro-Wilk test for normality of old premenopausal patients'
survival data: p =", shapiro_test_oldpremenopausal$p.value, "\n")
                cat("Shapiro-Wilk test for normality of young postmenopausal
patients' survival data: p =", shapiro_test_youngpostmenopausal$p.value, "\n")
              # compare survival
                premenopausal_survival <- oldpremenopausal$Survival
                postmenopausal_survival <- youngpostmenopausal$Survival
              # perform t-test
                t_test_result <- t.test(premenopausal_survival,
                                        postmenopausal_survival,
                                        alternative = "less")
              # print results of t-test
                cat("t-test for difference in survival between older premenopausal
and younger postmenopausal patients: p =", t_test_result$p.value, "\n")


        # UNIVARIATE

          # pre-menopausal group modeling of age and survival
            premenopausal_model <- coxph(Surv(Survival, Event) ~ Age,
                                         data = pre_meno)
            summary(premenopausal_model)

          # post-menopausal group modeling of age and survival
            postmenopausal_model <- coxph(Surv(Survival, Event) ~ Age,
                                          data = post_meno)
            summary(postmenopausal_model)

          # Wilcox
            wilcox.test(pre_meno$Age ~ pre_meno$Event) # Wilcoxon rank-sum test for
age and survival
            wilcox.test(post_meno$Age ~ post_meno$Event) # Wilcoxon rank-sum test for
age and survival

          # visuals
            # KM curves curves for pre-menopausal and post-menopausal age groups
              km_pre <- survfit(Surv(Survival, Event) ~
                                  cut(Age, breaks=c(0, 40, 50, 60,
                                                    max(pre_meno$Age))),
                                data=pre_meno)
              km_post <- survfit(Surv(Survival, Event) ~
                                   cut(Age, breaks=c(0, 40, 50, 60,
                                                     max(post_meno$Age))),
                                 data=post_meno)
              # plot
                plot(km_pre, col=c("red","blue","green","orange"),
                     xlab="Time (days)", ylab="Survival Probability",
                     main="Kaplan-Meier Curves by Age Group - Pre-Menopausal")
                  # legend
                    legend("bottomleft",
                           legend = c("<40", "41-50", "51-60", ">60"),
                           col=c("red","blue","green","orange"), lty=1, cex=0.8)
              # plot
                plot(km_post, col=c("red","blue","green","orange"),
                     xlab="Time (days)", ylab="Survival Probability",
                     main="Kaplan-Meier Curves by Age Group - Post-Menopausal")
                  # legend
                    legend("bottomleft", legend = c("<40", "41-50", "51-60", ">60"),
                           col=c("red","blue","green","orange"), lty=1, cex=0.8)
              # plot both
                plot(km_pre, col=c("red","blue","green","orange"),
```

```
                     xlab="Time (days)", ylab="Survival Probability",
                     main="Kaplan-Meier Curves by Age Group - Pre-Menopausal")
              lines(km_post, col=c("red","blue","green","orange"),
                    lty=2) # dashed post meno lines
              legend("bottomleft", legend=c("<40", "41-50", "51-60", ">60"),
                     col=c("red","blue","green","orange"), lty=1)


              # log-rank test for pre-menopausal group
              pre_meno_test <- survdiff(Surv(Survival, Event) ~ cut
                                        (Age, breaks=c(0, 40, 50, 60,
                                                       max(pre_meno$Age))),
                                        data=pre_meno)
              pre_meno_test

              # log-rank test for post-menopausal group
              post_meno_test <- survdiff(Surv(Survival, Event) ~
                                           cut(Age, breaks=c(0, 40, 50, 60,
                                                             max(post_meno$Age))),
                                         data=post_meno)
              post_meno_test




      # MULTIVARIATE

          # Fit a multivariate Cox proportional hazards model
            cox_mod <- coxph(Surv(Survival, Event) ~ Age + MenopauseStatus,
                             data = data)
            # Output the results
            summary(cox_mod)

          # Interaction test of age and menopausal status on survival
            interaction_model <- coxph(Surv(Survival, Event) ~ Age*MenopauseStatus,
                                       data = data)
            summary(interaction_model)

          # linear regression with interaction term
              summary(lm(Survival ~ Age + MenopauseStatus + age.menopause +
Event,
                         data = data))
            # Fit a multivariate regression model with interaction terms
              lin_reg_model <- lm(cbind(Survival, Event) ~ Age * MenopauseStatus,
                                  data = data)
             # Extract the model summary
                summary(lin_reg_model)


# QUESTION 4

      # part a

            # load both datasets
              data1 <- read.table("assignOct3-1.txt", header = TRUE)
              data2 <- read.table("assignOct4.txt", header = TRUE)

            # join datasets by patient IDs
              merged_data <- merge(data1, data2, by = "Ptid")

            # define good and poor survival groups
              good_survival <- merged_data[merged_data$Survival > 5 * 365, ]
              poor_survival <- merged_data[merged_data$Survival < 365, ]

            # define gene expression cols in data
              gene_cols <- c("Gene1", "Gene2", "Gene3", "Gene4", "Gene5")

            # wilcoxon rank sum for each gene
            for (i in gene_cols) {
```

```
                col_index <- which(names(merged_data) == i) # find column index of
gene in  merged dataset
                wilcox_result <- wilcox.test(good_survival[, col_index],
                                             poor_survival[, col_index]) # perform
Wilcoxon rank sum test
                p_value_wilcox <- wilcox_result$p.value

                if (p_value_wilcox < 0.05) {
                  print(paste(i, "has a significant difference in expression levels
(p =", p_value_wilcox, ")"))
                } else {
                  print(paste(i, "does not have a significant difference in
expression levels (p =", p_value_wilcox, ")"))
                }
              }
                    # apply Bonferroni correction
                    bonferroni_threshold <- 0.05/length(gene_cols)

                    # wilcoxon rank sum for each gene
                    for (i in gene_cols) {
                      col_index <- which(names(merged_data) == i) # find column
index of gene in merged dataset
                      wilcox_result <- wilcox.test(good_survival[, col_index],
                                                   poor_survival[, col_index]) #
perform Wilcoxon rank sum test
                      p_value_wilcox <- wilcox_result$p.value

                      if (p_value_wilcox < bonferroni_threshold) {
                        print(paste(i, "has a significant difference in expression
levels (p =", p_value_wilcox, ")"))
                      } else {
                        print(paste(i, "does not have a significant difference in
expression levels (p =", p_value_wilcox, ")"))
                      }
                    }


                    # new column for survival group
                    merged_data$SurvivalGroup <- ifelse(merged_data$Survival > 5
* 365, "good", ifelse(merged_data$Survival < 365, "poor", NA))

                    # ANOVA for each gene with post hoc Tukey HSD
                    for (i in gene_cols) {
                      col_index <- which(names(merged_data) == i) # find column
index of gene in merged dataset
                      anova_result <- aov(as.formula(paste(i, " ~
SurvivalGroup")), data = merged_data) # perform ANOVA
                      p_value_anova <- summary(anova_result)[[1]][["Pr(>F)"]][1]
# extract p-value from ANOVA summary

                      if (p_value_anova < 0.05) {
                        print(paste(i, "has a significant difference in
expression levels (p =", p_value_anova, ")"))
                        tukey_result <- TukeyHSD(anova_result) # perform Tukey
test
                        print(tukey_result)
                      } else {
                        print(paste(i, "does not have a significant difference in
expression levels (p =", p_value_anova, ")"))
                      }
                    }

              # box plots
                library(ggplot2)
                for (i in gene_cols) { #loop genes
                  col_index <- which(names(merged_data) == i) # find gene cols
                  plot_data <- data.frame(   # create data frame
                    expression = c(good_survival[, col_index],
                                   poor_survival[, col_index]),
```

```
                    survival = rep(c("Good", "Poor"), c(nrow(good_survival),
                                                        nrow(poor_survival)))
                  )
                  boxplots <- ggplot(plot_data, aes(x = survival,
                                                    y = expression,
                                                    fill = survival)) + # create box
plots
                    geom_boxplot() +
                    labs(x = "Survival", y = i) +  # axis labels
                    scale_fill_manual(values = c("Good" = "green", "Poor" = "red")) +
# colours
                    theme_bw()
                  print(boxplots)
                }


          # part b

                  # cox models
                    # Gene1
                      cox_model_gene1 <- coxph(Surv(Survival, Event) ~ Gene1,
                                               data=merged_data)
                      summary(cox_model_gene1)
                    # Gene2
                      cox_model_gene2 <- coxph(Surv(Survival, Event) ~ Gene2,
                                               data=merged_data)
                      summary(cox_model_gene2)
                    # Gene3
                      cox_model_gene3 <- coxph(Surv(Survival, Event) ~ Gene3,
                                               data=merged_data)
                      summary(cox_model_gene3)
                    # Gene4
                      cox_model_gene4 <- coxph(Surv(Survival, Event) ~ Gene4,
                                               data=merged_data)
                      summary(cox_model_gene4)
                    # Gene5
                      cox_model_gene5 <- coxph(Surv(Survival, Event) ~ Gene5,
                                               data=merged_data)
                      summary(cox_model_gene5)

                  # visuals
                    # boxplots
                      # load  library
                        library(ggplot2)
                      # vector with gene names to plot
                        gene_names <- paste0("Gene", 1:5)
                      # loop
                      for (gene in gene_names) {
                        # df of the expression values
                        # sep by  event
                        plot_data <- data.frame(
                          expression = c(merged_data[merged_data$Event == 0, gene],
                                         merged_data[merged_data$Event == 1, gene]),
                          survival = rep(c("Alive", "Dead"), c(sum(merged_data$Event ==
0),
                                                               sum(merged_data$Event ==
1)))
                        )
                        # boxplot for current gene
                        boxplots <- ggplot(plot_data, aes(x = survival, y = expression,
fill = survival)) +
                          geom_boxplot() +
                          labs(x = "Event (0 = Alive, 1 = Dead)", y = gene) + # label
the axes
                          scale_fill_manual(values = c("Alive" = "green", "Dead" =
"red")) + # set colours
                          theme_bw() # set plot theme
                        print(boxplots) # print plot
                      }
```