Identify Biomarkers for Cervical Cancer

November, 2020

1. Abstract

To identify biomarkers for cervical cancer, we capture the relationship between biomarkers and the disease free survival of patients by cox regression model, adjust the impact of other covariates, and select significant inputs by stepwise cox regression.

2. Data Description

After data wrangling, we now have 6 biomarkers, 7 continuous covariates and 16 categorical covariates with 287 observations. Here is a quick summary of the variables.

Biomarkers

```
summary(biomarkers)
```

```
##      UBE2C            PDL1          HPV18           HPV58      
##  Min.   :  0.00   Min.   :  0.0   Min.   :  0.0   Min.   : 10.0  
##  1st Qu.:  5.50   1st Qu.: 70.0   1st Qu.:100.0   1st Qu.:120.0  
##  Median : 20.00   Median :120.0   Median :150.0   Median :160.0  
##  Mean   : 54.22   Mean   :124.7   Mean   :152.3   Mean   :161.8  
##  3rd Qu.: 90.00   3rd Qu.:175.0   3rd Qu.:190.0   3rd Qu.:210.0  
##  Max.   :285.00   Max.   :300.0   Max.   :300.0   Max.   :277.5  
##  NA's   :16       NA's   :16      NA's   :16      NA's   :16     
##      HPV16           ASCC2       
##  Min.   :  5.00   Min.   :  0.00  
##  1st Qu.: 95.75   1st Qu.: 47.50  
##  Median :202.50   Median : 77.50  
##  Mean   :184.80   Mean   : 89.04  
##  3rd Qu.:270.00   3rd Qu.:115.25  
##  Max.   :300.00   Max.   :300.00  
##  NA's   :16       NA's   :16      
```

Continuous Covariates

```
summary(continuous_inputs)
```

```
##       age            height        weight           BMI          tumor.size    
##  Min.   :25.00   Min.   :135   Min.   :29.00   Min.   :13.70   Min.   :  1.00  
##  1st Qu.:46.00   1st Qu.:151   1st Qu.:50.50   1st Qu.:21.21   1st Qu.: 25.00  
##  Median :55.00   Median :155   Median :57.00   Median :23.28   Median : 38.00  
##  Mean   :55.39   Mean   :155   Mean   :57.78   Mean   :24.02   Mean   : 40.66  
##  3rd Qu.:64.00   3rd Qu.:159   3rd Qu.:64.00   3rd Qu.:26.70   3rd Qu.: 50.00  
##  Max.   :98.00   Max.   :171   Max.   :92.50   Max.   :41.66   Max.   :110.00  
##                  NA's   :6     NA's   :6       NA's   :6       NA's   :27      
##       CEA                SCC        
##  Min.   :   0.500   Min.   : 0.100  
##  1st Qu.:   1.500   1st Qu.: 0.600  
##  Median :   2.450   Median : 1.500  
##  Mean   :  11.969   Mean   : 6.476  
##  3rd Qu.:   4.775   3rd Qu.: 5.600  
##  Max.   :1396.000   Max.   :63.700  
##  NA's   :69         NA's   :50      
```

Categorical Covariates

```
summary(categorical_inputs)
```

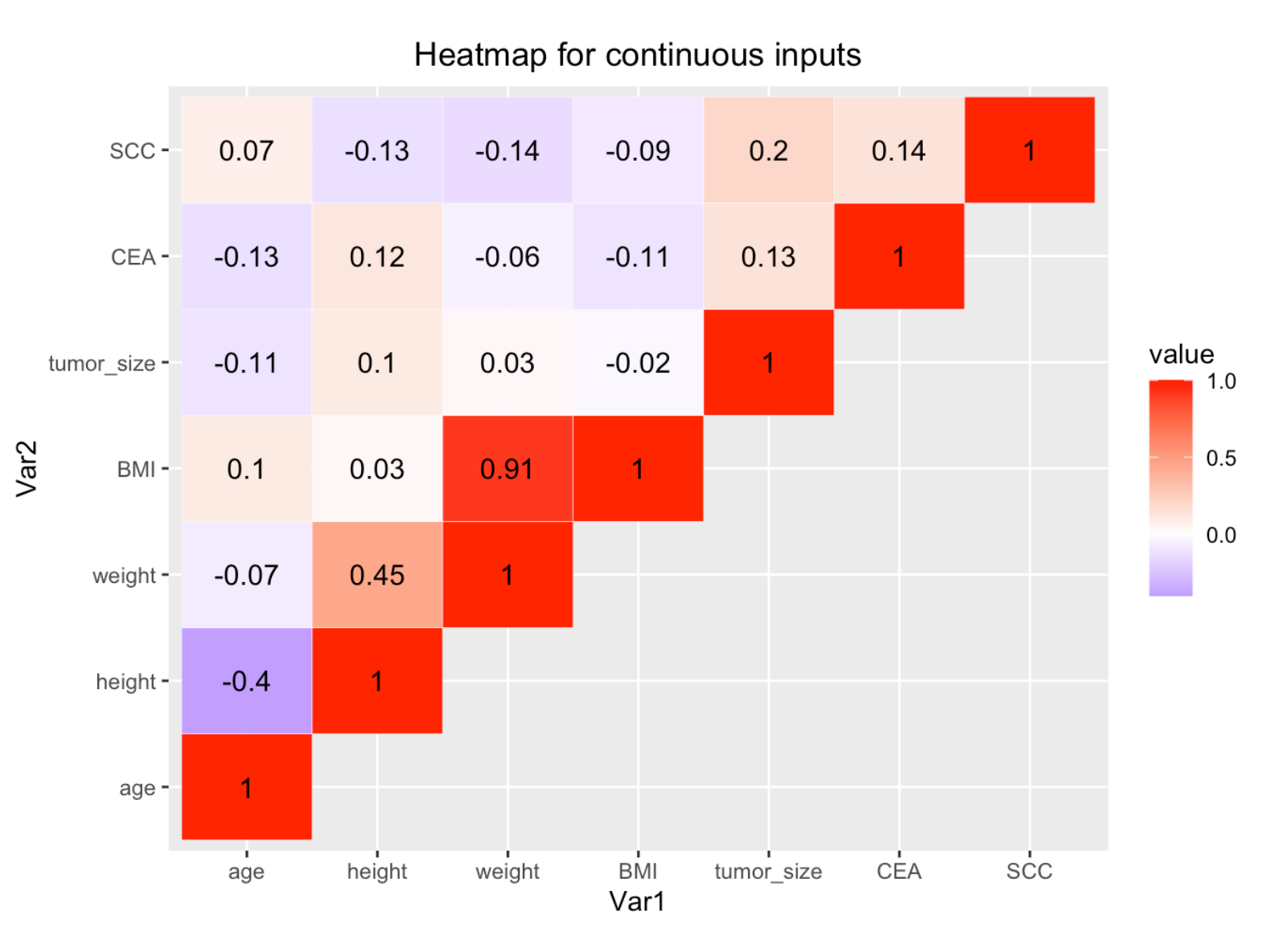
```
##  Marriage         G            P           A       menopause  high.blood.pressure
##  0   : 37   4      :84   2      :84   0   :134   0   : 90   0   :170          
##  1   :248   3      :62   3      :75   1   : 66   1   :145   1   : 83          
##  NA's:  2   2      :34   4      :45   2   : 52   NA's: 52   NA's: 34          
##             5      :27   0      :19   3   : 18                                
##             6      :23   5      :18   4   :  3                                
##             (Other):48   (Other):37   6   :  1                                
##             NA's   : 9   NA's   : 9   NA's: 13                                
##  diabetes   grading      c.stage       p.stage     FIGO.Stage RT      CT     
##  0   :200   1   : 15   1B1    :105   1B1    :75   IB1    :95   0:118   0:168  
##  1   : 42   2   :193   3B     : 47   3B     :40   IIB    :43   1:169   1:119  
##  NA's: 45   3   : 45   1B2    : 31   2B     :27   IB2    :41                  
##             4   :  5   2B     : 20   1B2    :25   IIIB   :23                  
##             NA's: 29   4B     : 19   1A1    :14   IB     :16                  
##                        (Other): 55   (Other):39   (Other):67                  
##                        NA's   : 10   NA's   :67   NA's   : 2                  
##   smoke     betel.nut   alcohol   
##  0   :223   0   :232   0   :222  
##  1   : 13   1   :  1   1   : 14  
##  NA's: 51   NA's: 54   NA's: 51  
## 
## 
## 
## 
```

3. Screening Out Noisy Covariates

We try to exclude some inappropate covariates by correlation heatmap, univariate cox regression and log-rank test.

Pearson Correlation Heatmap for Continuous Covariates

First, we check the correlation between continuous covariates to avoid highly-correlated inputs. The Pearson correlation heatmap for the 7 continuous covariates is as follows. We discover that there is a high correlation (corr = 0.91) between BMI and weight.



We further check that this correlation statistics is significantly different from 0, with p-value < 2.2e-16. Therefore, we can exclude one of them from our inputs (we randomly choose "BMI").

```
#test whether the corr between weight and BMI is significantly different from 0
cor.test(cts_input$weight, cts_input$BMI, method="pearson", use="complete.obs")
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  cts_input$weight and cts_input$BMI
## t = 35.615, df = 279, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8817735 0.9244479
## sample estimates:
##       cor 
## 0.9053711
```

Univariate Cox Regression for Continuous Covariates

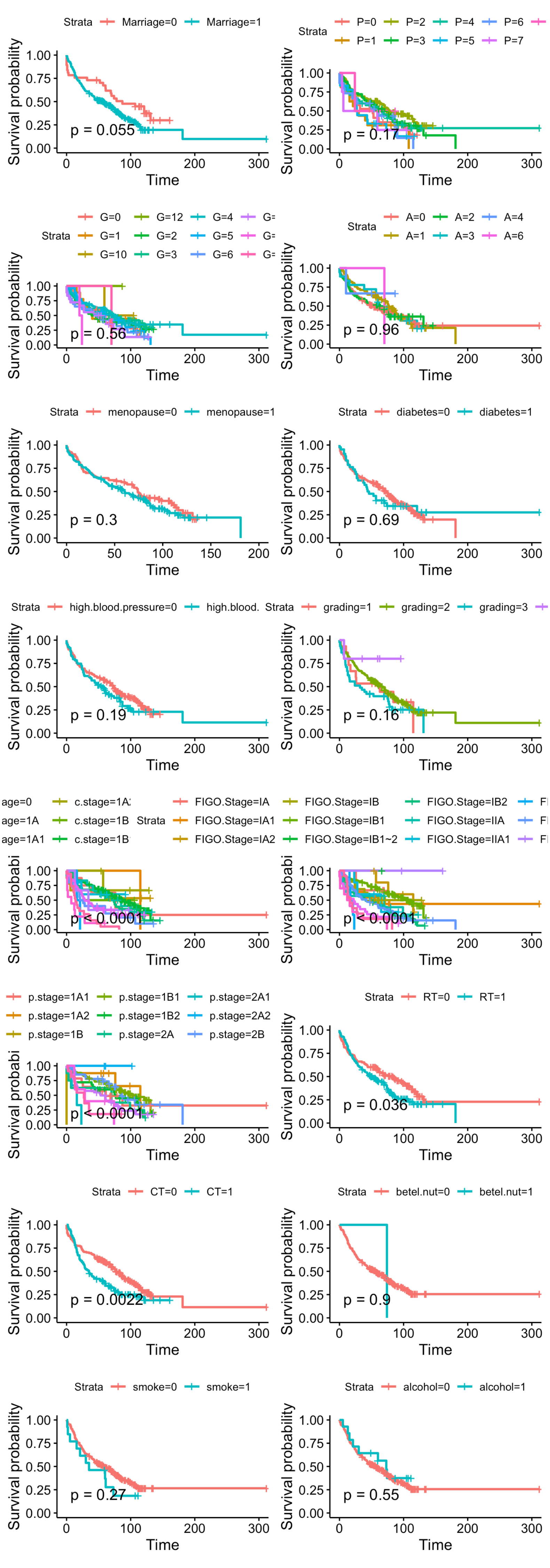
We then conduct univariate cox regression to check the significance of each variable. Since this is only a preliminary screening, we use a looser criterion (p-value < 0.1) to determine whether to retain this variable or not.

```
## [1] "age: p-value=0.012886"
## [1] "height: p-value=0.972536"
## [1] "weight: p-value=0.056573"
## [1] "tumor_size: p-value=0.000041"
## [1] "CEA: p-value=0.457277"
## [1] "SCC: p-value=0.000106"
```

From the result, we decide to exclude "height" and "CEA" from our inputs.

Kaplain-Meier Curve for Categorical Covariates

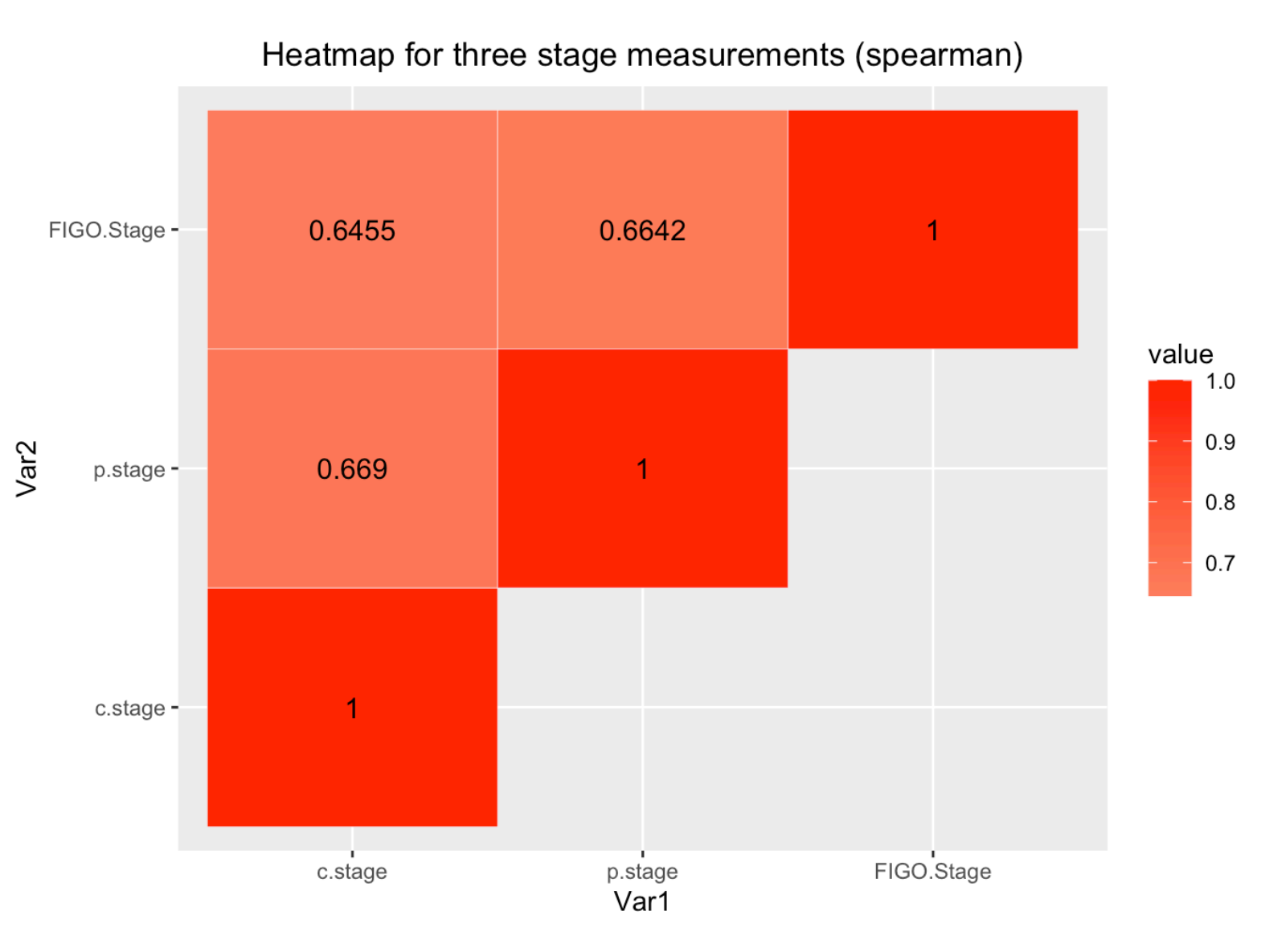
For each categorical covariate, we plot the Kaplan-Meier curve with the p-value of log-rank test printing in the lower left corner to check whether each category is significantly apart. Similarly, we use a loose criterion (p-value < 0.1) to determine whether to retain this variable or not.



From the results, the variables with p-value < 0.1 are: "marriage", "c stage", "p stage", "FIGO stage", "RT", and "CT", and we exclude all the other categorical variables.

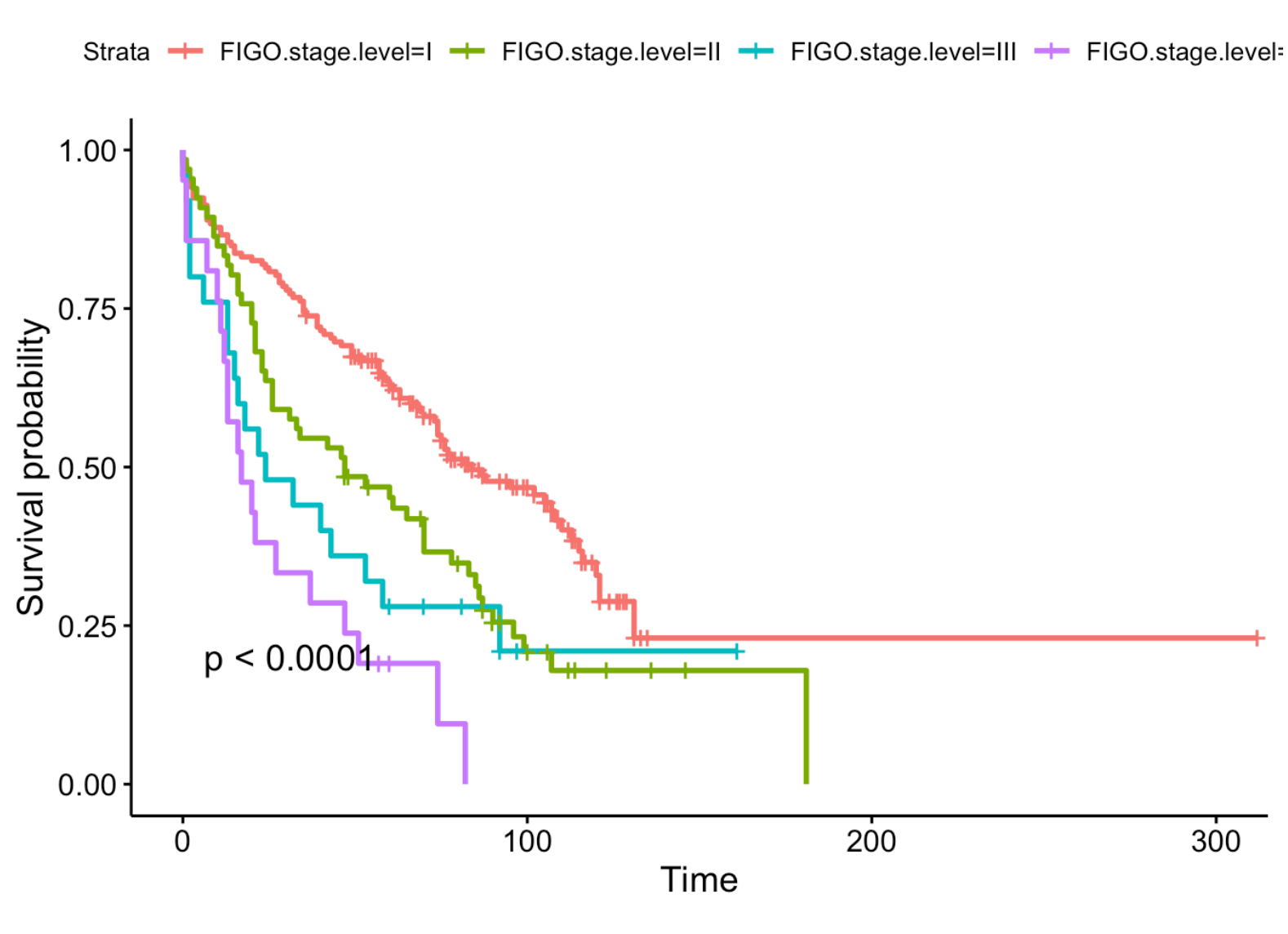
Spearman Correlation Heatmap for Ordinal Categorical Covariates

We now check the correlation between "c stage", "p stage" and "FIGO stage" by plotting the Spearman correlation heatmap. We can discover that the correlation between any two of them is high (>0.6). We further check that all the three correlation statistics are significantly different from 0, with p-value < 2.2e-16. Therefore, we can exclude any two of them from our inputs. Since "FIGO stage" has the least number of NAs, we choose to retain "FIGO stage".



Merging Groups for FIGO Stage

We first try to merge the FIGO stage variable with level: I, II, III, and IV. The Kaplan-Meier curve and the result of pairwise log-rank test are as follows.

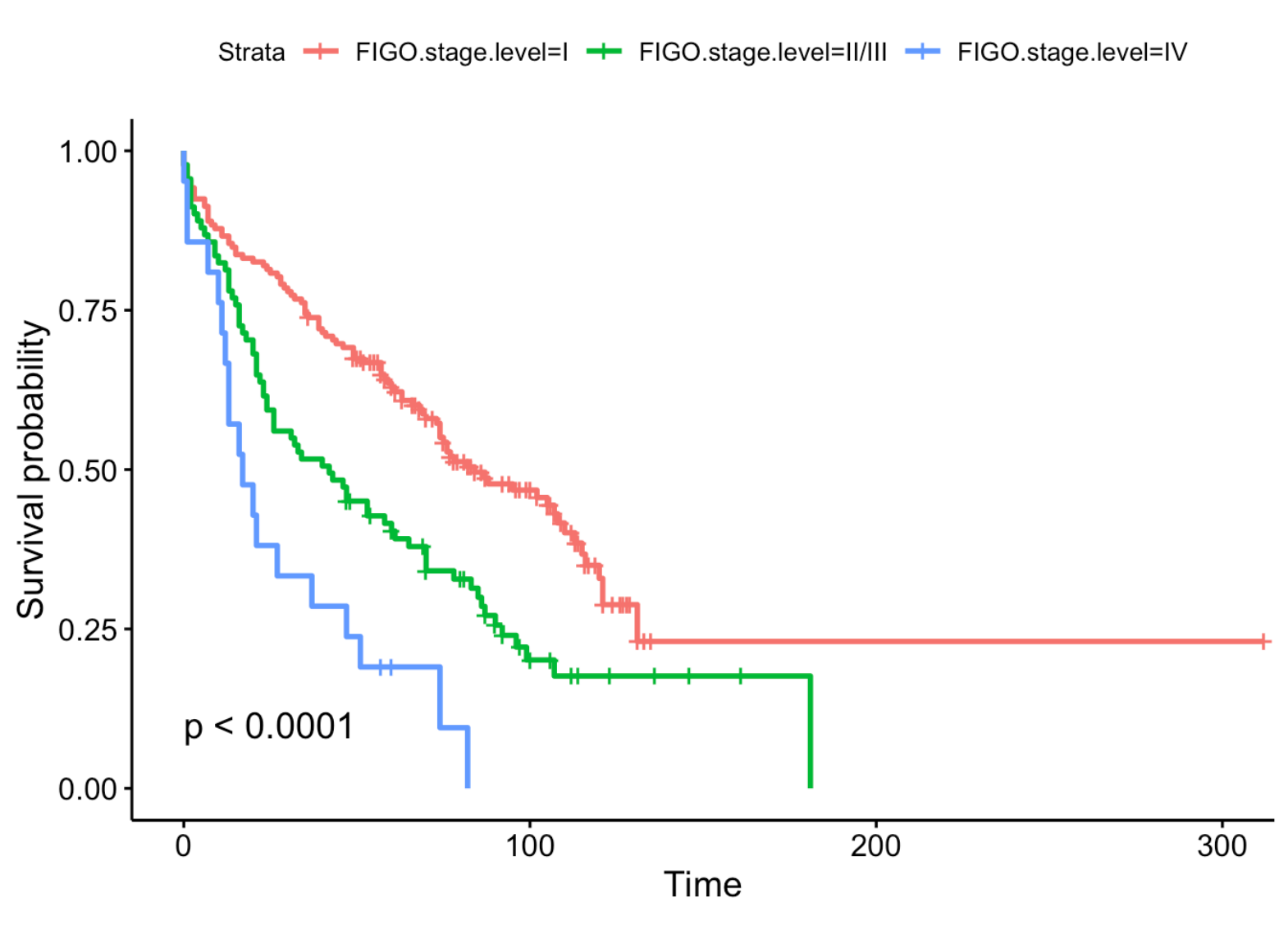


Pairwise Log-Rank Test for Level: I, II, III, IV

```
##          I      II     III   
## II  0.0037     NA     NA    
## III 0.0037 0.3161     NA    
## IV  0.0000 0.0037 0.2038
```

Although the overall log-rank test seems signifacant, with p-value < 0.0001, the pairwise log-rank test reveals that level II and level III are not significantly apart. As a result, we try to further merge level II and level III together.

The Kaplan-Meier curve of FIGO stage with level: I, II/III, IV and the result of pairwise log-rank test are as follows.



Pairwise Log-Rank Test for Level: I, II/III, IV

```
##            I II/III
## II/III 2e-04     NA
## IV     0e+00 0.0065
```

After merging level II and III as a group, we now have a nice result from the pairwise log-rank test, with each level significantly different from another. Therefore, we will stick to this grouping in the following analysis.

The Remaining Covariates

After screening out noisy covariates, the remaining covariates and the six biomarkers are:

- Continuous covariates: age, weight, tumor size, SCC
- Categorical covariates: marriage, RT, CT, FIGO stage
- Biomarkers: UBE2C, PDL1, HPV18, HPV58, HPV16, ASCC2

These are the variables that we will use in our following cox regression.

4. Finding cutpoints for Biomarkers & Stepwise Cox Regression

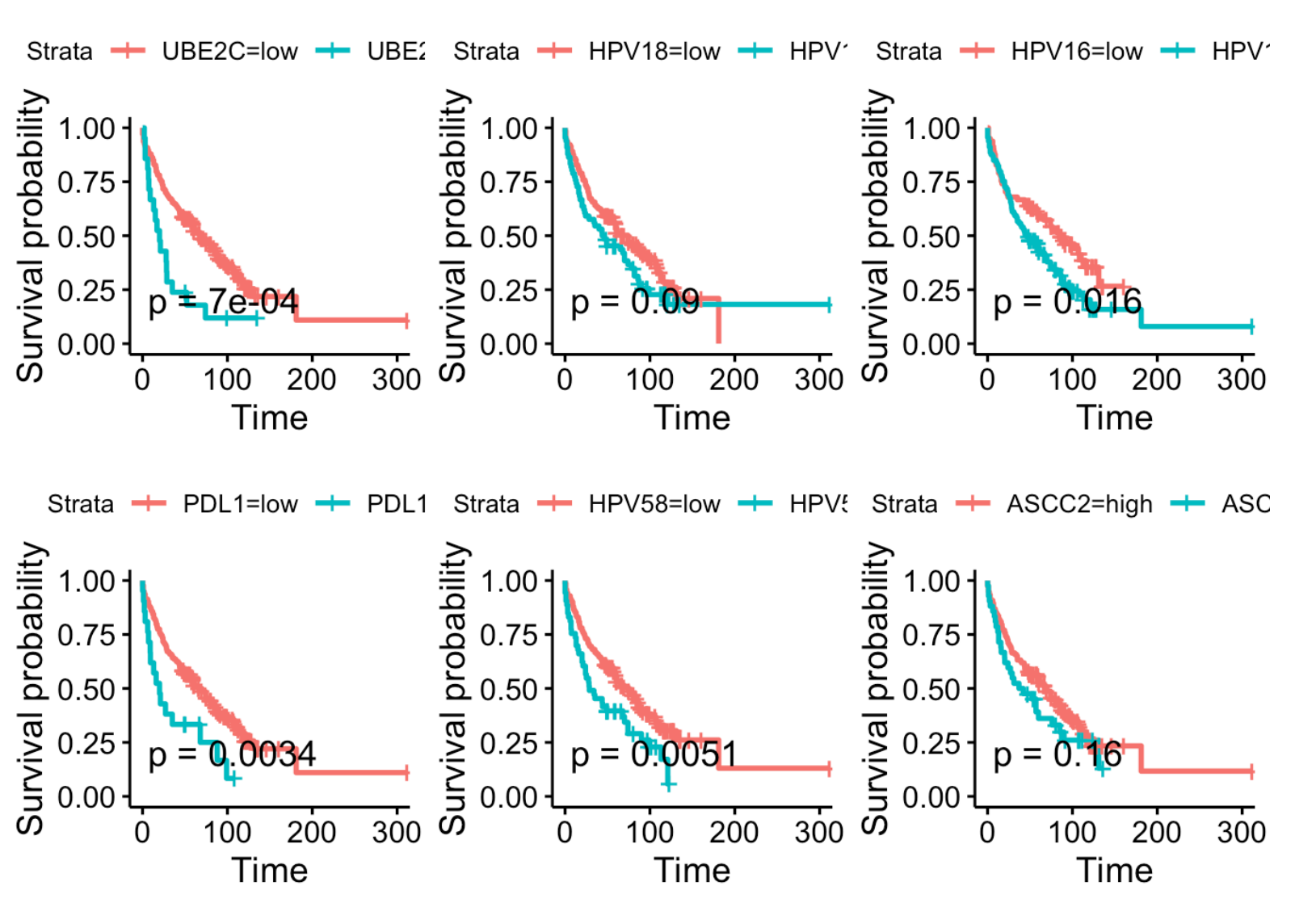
First, we subset our data and focus on patients with "squamous cell carcinoma". We then find 2~3 cutpoints for each of the six biomarkers by `surv_cutpoint()` function in `survminer` package or `rhier()` function in `rolr` package.

One Optimal Cutpoint by `surv_cutpoint()`

The cutpoints for the six biomarkers are:

```
##       cutpoint statistic
## UBE2C    135.0  2.707632
## PDL1     200.0  2.372516
## HPV18    175.0  1.657188
## HPV58    200.0  2.592401
## HPV16    172.5  2.432752
## ASCC2     45.0  1.317476
```

We can now plot the Kaplan-Meier Curve for each biomarker to see whether the cutting is significant.



Finally, we can perform stepwise cox regression to conduct full model selection.

```
## Call:
## coxph(formula = Surv(time, status) ~ UBE2C + PDL1 + HPV18 + HPV16 + 
##     ASCC2 + age + weight + tumor_size + FIGO_stage, data = cox_reg_data)
## 
##   n= 151, number of events= 101 
## 
##                       coef exp(coef)  se(coef)      z Pr(>|z|)    
## UBE2Chigh         1.207530  3.345213  0.326108  3.703 0.000213 ***
## PDL1high          0.586004  1.796794  0.309289  1.895 0.058135 .  
## HPV18high         0.330544  1.391725  0.225835  1.464 0.143289    
## HPV16high         0.469650  1.599434  0.220471  2.130 0.033154 *  
## ASCC2low          0.658650  1.932182  0.263759  2.497 0.012519 *  
## age               0.032096  1.032617  0.009846  3.260 0.001115 ** 
## weight           -0.021973  0.978266  0.010949 -2.007 0.044755 *  
## tumor_size        0.022160  1.022407  0.005664  3.912 9.14e-05 ***
## FIGO_stageII/III -0.051403  0.949896  0.246868 -0.208 0.835056    
## FIGO_stageIV      0.890648  2.436707  0.365836  2.435 0.014910 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                  exp(coef) exp(-coef) lower .95 upper .95
## UBE2Chigh           3.3452     0.2989    1.7654    6.3388
## PDL1high            1.7968     0.5565    0.9800    3.2943
## HPV18high           1.3917     0.7185    0.8940    2.1666
## HPV16high           1.5994     0.6252    1.0382    2.4640
## ASCC2low            1.9322     0.5175    1.1522    3.2401
## age                 1.0326     0.9684    1.0129    1.0527
## weight              0.9783     1.0222    0.9575    0.9995
## tumor_size          1.0224     0.9781    1.0111    1.0338
## FIGO_stageII/III    0.9499     1.0527    0.5855    1.5410
## FIGO_stageIV        2.4367     0.4104    1.1896    4.9912
## 
## Concordance= 0.705  (se = 0.028 )
## Likelihood ratio test= 57.26  on 10 df,   p=1e-08
## Wald test            = 58.91  on 10 df,   p=6e-09
## Score (logrank) test = 63.78  on 10 df,   p=7e-10
```

From the regression result, the selected biomarkers and covariates are listed below. If the selected variable is not significant, its p-value is added in a parenthesis for your reference.

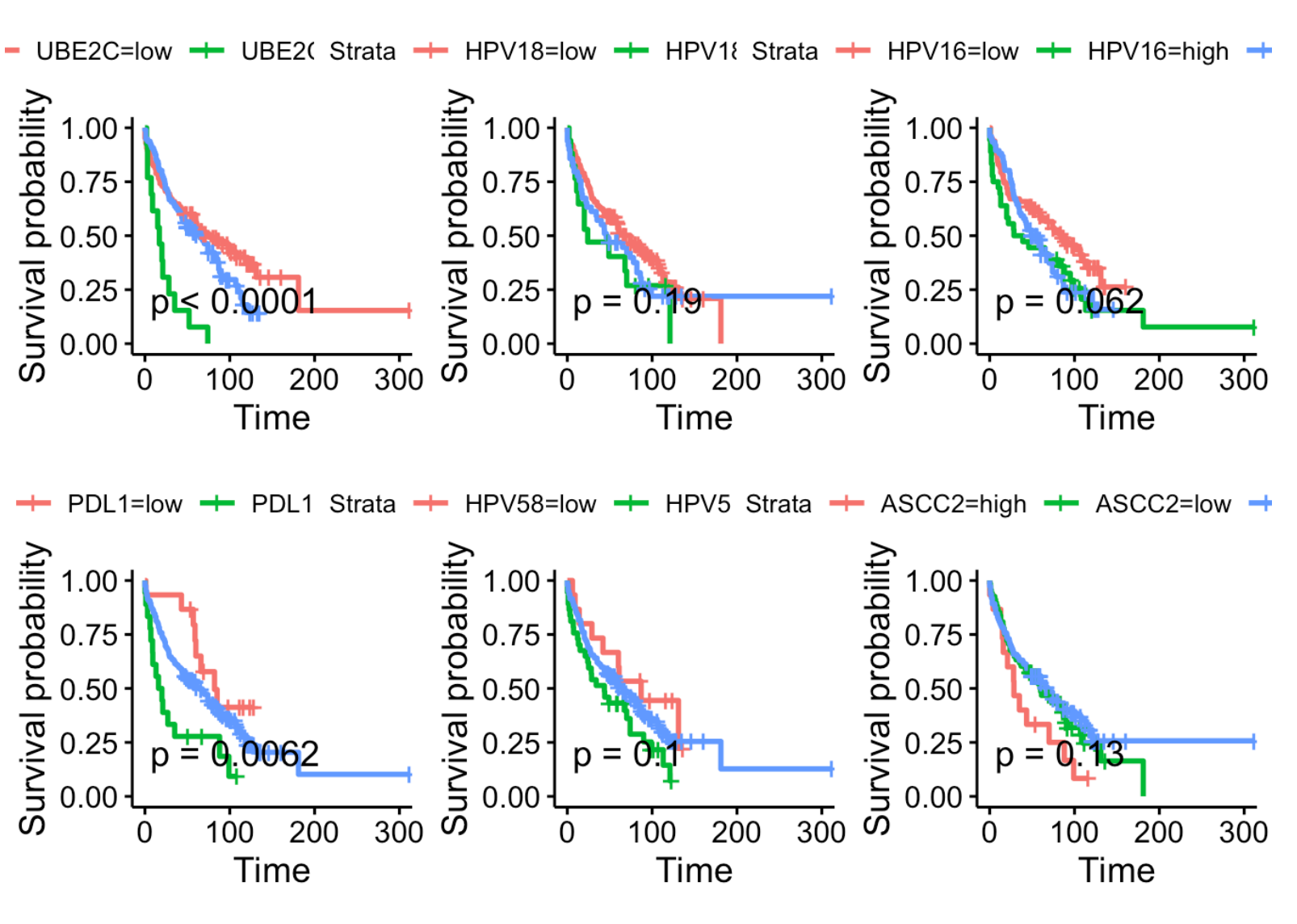
- Biomarkers: UBE2C, PDL1 (0.058), HPV18 (0.143), HPV16, ASCC2
- Continuous covariates: age, weight, tumor size
- Categorical covariates: FIGO stage

Two Optimal Cutpoints by `rhier()`

The cutpoints for the six biomarkers are:

```
## [1] "UBE2C"
## Best splits from hierarchical method : 20 160 
## 
## [1] "PDL1"
## Best splits from hierarchical method : 42.5 217.5 
## 
## [1] "HPV18"
## Best splits from hierarchical method : 180 240 
## 
## [1] "HPV58"
## Best splits from hierarchical method : 80 210 
## 
## [1] "HPV16"
## Best splits from hierarchical method : 180 270 
## 
## [1] "ASCC2"
## Best splits from hierarchical method : 65 172.5
```

We can now plot the Kaplan-Meier Curve for each biomarker to see whether the cutting is significant.



Finally, we can perform stepwise cox regression to conduct full model selection.

```
## Call:
## coxph(formula = Surv(time, status) ~ UBE2C + HPV16 + age + weight + 
##     tumor_size + FIGO_stage, data = cox_reg_data)
## 
##   n= 151, number of events= 101 
## 
##                       coef exp(coef)  se(coef)      z Pr(>|z|)    
## UBE2Chigh         1.489087  4.433044  0.394189  3.778 0.000158 ***
## UBE2Cmedium       0.094068  1.098635  0.226460  0.415 0.677859    
## HPV16high         0.891536  2.438872  0.283630  3.143 0.001671 ** 
## HPV16medium       0.279875  1.322965  0.229532  1.219 0.222719    
## age               0.034371  1.034968  0.010120  3.396 0.000683 ***
## weight           -0.023669  0.976609  0.010722 -2.207 0.027280 *  
## tumor_size        0.019255  1.019442  0.005792  3.324 0.000886 ***
## FIGO_stageII/III -0.089429  0.914454  0.246105 -0.363 0.716324    
## FIGO_stageIV      1.028474  2.796795  0.352807  2.915 0.003556 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##                  exp(coef) exp(-coef) lower .95 upper .95
## UBE2Chigh           4.4330     0.2256    2.0472    9.5992
## UBE2Cmedium         1.0986     0.9102    0.7048    1.7124
## HPV16high           2.4389     0.4100    1.3988    4.2522
## HPV16medium         1.3230     0.7559    0.8437    2.0746
## age                 1.0350     0.9662    1.0146    1.0557
## weight              0.9766     1.0240    0.9563    0.9973
## tumor_size          1.0194     0.9809    1.0079    1.0311
## FIGO_stageII/III    0.9145     1.0935    0.5645    1.4813
## FIGO_stageIV        2.7968     0.3576    1.4007    5.5843
## 
## Concordance= 0.673  (se = 0.029 )
## Likelihood ratio test= 49.37  on 9 df,   p=1e-07
## Wald test            = 52.94  on 9 df,   p=3e-08
## Score (logrank) test = 58.8  on 9 df,   p=2e-09
```

From the regression result, the selected biomarkers and covariates are:

- Biomarkers: UBE2C, HPV16
- Continuous covariates: age, weight, tumor size
- Categorical covariates: FIGO stage