DATA SCIENCE LAB IN MEDICINE

Big Data in Healthcare – Exam Project

Università degli Studi di Milano – Bicocca

Students:

Induni Sandapiumi Nawarathna Pitiyage – 906451

Sara Campolattano – 906453

Course Instructors:

Prof. Maria Grazia Valsecchi

Dr. Davide Bernasconi

Dr. Giulia Capitoli

TABLE OF CONTENTS

1. INTRODUCTION & DATASET DESCRIPTION

Breast cancer represents a significant public health concern, particularly due to its potential to metastasize, impacting women survival and quality of life.
This project aims to analyze a dataset derived from an observational study focusing on survival free from metastasis in 144 women diagnosed with breast cancer who have lymph node involvement.
This study endeavors to derive insights and predictive models that can enhance our understanding and management of breast cancer progression.

The dataset encompasses a range of clinical and gene expression variables, providing a comprehensive overview of factors that may influence patient outcomes in terms of their response to treatment.
The variables contained in the dataset include follow-up time (recorded in months), event indicator (binary variable where metastasis or death = 1, and censoring = 0), tumor features such as diameter (expressed with two levels, i.e., <= 2cm or >2cm), number of lymph nodes involved (expressed with two levels, i.e., < 4 or >= 4), estrogen receptor status (expressed with two levels, i.e., positive or negative), tumor grading (expressed with three ordered levels, i.e., poorly differentiated, intermediate, well differentiated), age of the patient at diagnosis (expressed in years), and the expression levels of 70 potentially prognostic genes.

The analyses carried out in this project are structured to achieve several objectives:

- **Descriptive Analysis**, to provide an overview of all variables in the dataset to understand their characteristics inherent patterns.
- **Univariate Analysis**, to examine the association of each clinical variable with the outcome (metastasis/death, or censoring) using the Cox proportional hazards model.
- **Basic Predictive Model Development**, to estimate the risk of metastasis or death using clinical variables.
- **Model Assumption Evaluation**, to assess the functional form of continuous variables and verify the proportional hazards assumption for all covariates.

- **Gene Expression Analysis**, to investigate the association of each gene expression variable with the outcome using univariate Cox models, followed by adjustments for multiple testing using the Benjamini-Hochberg method.
- **Penalized Cox Model with LASSO**, selecting gene expression variables significantly associated with the outcome through penalized regression techniques.
- **Augmented Predictive Model Development**, to integrate selected gene expression variables with clinical variables to develop a comprehensive predictive model.
- **Risk Prediction**, to predict the risk of the event at a fixed time-point (e.g., 1 year) for selected subjects using both the basic and augmented models.

2. DESCRIPTIVE ANALYSIS on clinical variables

To provide a comprehensive overview of clinical characteristics and follow-up outcomes of women affected by breast cancer with lymph node involvement, a descriptive analysis was carried out.

In *Table 1*, a summary depicting descriptive statistics of the clinical covariates can be seen.

```
##     Diam       N              ER              Grade         Age
##  <=2cm:73   <4 :106   Negative: 27   Intermediate:55   Min.   :26.00
##  >2cm :71   >=4: 38   Positive:117   Poorly diff :48   1st Qu.:41.00
##                                      Well diff   :41   Median :45.00
##                                                        Mean   :44.31
##                                                        3rd Qu.:49.00
##                                                        Max.   :53.00
##       time                          event    age_category
##  Min.   : 0.05476   Censoring          :96   26-35: 7
##  1st Qu.: 4.70568   Metastasis or Death:48   35-44:63
##  Median : 6.99521                            44-53:74
##  Mean   : 7.35130
##  3rd Qu.: 9.98631
##  Max.   :17.65914
##                time_var
##  FirstHalf_followUp :96
##  SecondHalf_followUp:48
##
##
##
##
```

Table 1

Total n. of patients: 144

From *Table 1*, it is possible to notice that **patients' age** at diagnosis ranges from 26 to 53 years, with a median age of 45 years. The interquartile range (41 to 49 years) indicates that most patients were diagnosed in their 40s, a common age range for breast cancer occurrence.

As for tumor-related features, it is possible to say that patients' **tumor size** is evenly distributed with 73 women having tumors $\leq$ 2 cm and 71 having tumors > 2 cm, offering a balanced perspective on how tumor diameter may influence outcomes. Most women (106) have fewer than 4 involved **lymph nodes**, suggesting that most patients in the study might be at a relatively earlier stage of lymph node involvement and that they have a potentially better prognosis compared to the 38 women with more extensive lymph node involvement. The cohort is predominantly **estrogen receptor**-positive (117 women), which is associated with better responsiveness to hormonal therapies and a favorable prognosis. **Tumor grade** distribution varies, with 55 (38%)

intermediate, 48 (33%) poorly differentiated, and 41 (28%) well-differentiated, highlighting cohort heterogeneity and varying prognostic implications.

We care to notice that **follow-up** times vary from 0.05 to approximately 18 months, with a median of 7 months. Out of 144 patients, 96 women were censored (alive/metastasis-free or lost to follow-up), while 48 experienced metastasis or death, indicating a relatively favorable short-term prognosis for the majority.

When categorized by age, 7 women are in the 26-35 range, 63 in 35-44, and 74 in 44-53, reflecting typical breast cancer patient distribution. The follow-up data, with 96 women at the beginning and 48 at the end, underscores the importance of temporal changes in understanding disease progression and survival outcomes.

In *Table 2*, an assessment of follow-up outcomes, specifically distinguishing between patients who were censored and those who experienced metastasis or death, can be found.

```
##
##                     FirstHalf_followUp SecondHalf_followUp Sum
##   Censoring                         55                 41  96
##   Metastasis or Death               41                  7  48
##   Sum                               96                 48 144
```

Table 2

Follow-up Time: 17.65914 months

First half of Follow-up Time: < 8.8 months

Second half of Follow-up Time: > 8.8 months

From *Table 2* it is possible to see that, during the first half (i.e., 8.8 months) of the follow-up period, 55 women were censored, meaning they were alive and metastasis-free or lost to follow-up. By the end (second half, i.e., after the first 8.8 months) of the follow-up period, this number decreased to 41.

During the first half of the follow-up period, 41 women experienced metastasis or death. By the end, this number decreased to 7.

In *Table 3*, we provide a comprehensive summary of the clinical covariates in relation to the tumor grading.

```
## datai$Grade: Intermediate
##     Diam      N             ER              Grade         Age
##  <=2cm:30   <4 :41   Negative: 5   Intermediate:55   Min.   :35.00
##  >2cm :25   >=4:14   Positive:50   Poorly diff : 0   1st Qu.:42.00
##                                    Well diff   : 0   Median :44.00
##                                                      Mean   :44.47
##                                                      3rd Qu.:48.50
##                                                      Max.   :52.00
##       time                           event    age_category
##  Min.   : 0.05476   Censoring          :36   26-35: 1
##  1st Qu.: 5.04449   Metastasis or Death:19   35-44:28
##  Median : 7.12115                            44-53:26
##  Mean   : 7.45024
##  3rd Qu.:10.37509
##  Max.   :17.24025
##               time_var
##  FirstHalf_followUp :37
##  SecondHalf_followUp:18
##
##
##
##
## ------------------------------------------------------------
## datai$Grade: Poorly diff
##     Diam      N             ER              Grade         Age
##  <=2cm:14   <4 :30   Negative:20   Intermediate: 0   Min.   :26.00
##  >2cm :34   >=4:18   Positive:28   Poorly diff :48   1st Qu.:41.00
##                                    Well diff   : 0   Median :45.00
##                                                      Mean   :44.21
##                                                      3rd Qu.:49.00
##                                                      Max.   :53.00
##       time                           event    age_category
##  Min.   : 0.3532   Censoring          :26   26-35: 4
##  1st Qu.: 2.3299   Metastasis or Death:22   35-44:18
##  Median : 6.6585                            44-53:26
##  Mean   : 6.7375
##  3rd Qu.: 9.4716
##  Max.   :17.6591
##               time_var
##  FirstHalf_followUp :32
##  SecondHalf_followUp:16
##
##
##
##
## ------------------------------------------------------------
## datai$Grade: Well diff
##     Diam      N             ER              Grade         Age
##  <=2cm:29   <4 :35   Negative: 2   Intermediate: 0   Min.   :29.0
##  >2cm :12   >=4: 6   Positive:39   Poorly diff : 0   1st Qu.:41.0
##                                    Well diff   :41   Median :45.0
##                                                      Mean   :44.2
##                                                      3rd Qu.:49.0
##                                                      Max.   :52.0
##       time                          event    age_category
##  Min.   : 2.335   Censoring          :34   26-35: 2
##  1st Qu.: 5.574   Metastasis or Death: 7   35-44:17
##  Median : 7.570                            44-53:22
##  Mean   : 7.937
##  3rd Qu.: 9.999
##  Max.   :16.148
##               time_var
##  FirstHalf_followUp :27
##  SecondHalf_followUp:14
##
##
##
##
```

Table 3

From the above table, it is possible to see that patients with intermediate-grade tumor show an approximately balanced distribution in tumor size, predominantly positive estrogen receptor status, and most have fewer than 4 involved lymph nodes. The age range is relatively narrow, with most patients being in their 40s. Follow-up time data indicates a higher proportion of censored cases (65%), suggesting a relatively favorable prognosis within this group.

On the other hand, patients with poorly differentiated tumors show larger tumor sizes and a higher proportion of estrogen receptor-negative cases compared to intermediate-grade tumors. This group also has a higher rate of adverse events (46% experiencing metastasis or death), indicating a poorer prognosis. The age distribution is slightly wider, but most patients are in their 40s and early 50s.

As for patients with well-differentiated tumors, they seem to predominantly have smaller tumors, positive estrogen receptor status, and fewer involved lymph nodes. The follow-up data shows the lowest rate of adverse events (17%), suggesting the best prognosis among the three grades. The age distribution is mostly similar to the other groups, with most patients in their 40s.

In *Figure 1* and in Table 4, the plot of the Kaplan-Meier survival curve and the summary can be found.

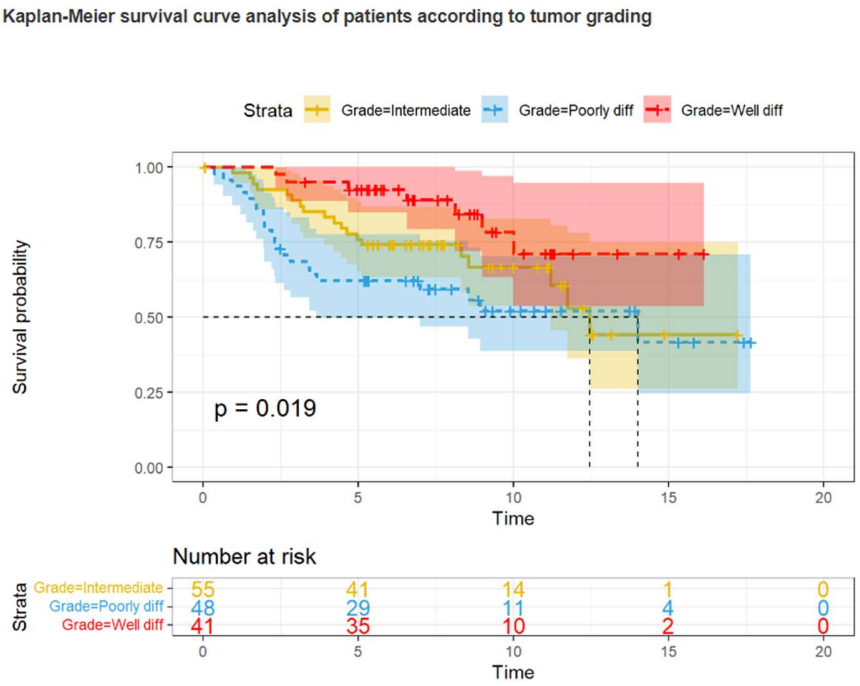

Figure 1

```
##                    records n.max n.start events    rmean se(rmean)   median
## Grade=Intermediate      55    55      55     19 11.93223  1.008870 12.46543
## Grade=Poorly diff       48    48      48     22 10.42813  1.112158 14.01232
## Grade=Well diff         41    41      41      7 14.68470  0.996644       NA
##                    0.95LCL 0.95UCL
## Grade=Intermediate 11.211499      NA
## Grade=Poorly diff   3.655031      NA
## Grade=Well diff           NA      NA
```

Table 4

As it can be seen from the Kaplan-Meier Survival Curve plot shown in Figure 1, patients with **well-differentiated** tumors have the highest survival probability over time, indicating the best prognosis among the three groups. The survival curve shows a relatively shallow decline, suggesting fewer adverse events.

On the other hand, patients with **intermediate-grade** tumors have a moderate survival probability, with the survival curve showing a more pronounced decline compared to the well-differentiated group.

As one can expect patients with **poorly differentiated** tumors have the lowest survival probability. The curve shows a steep decline, indicating a higher rate of adverse events such as metastasis or death.

As the **p-value** is 0.019, it is possible to say that the differences in survival probabilities among the three tumor grades are statistically significant, suggesting that tumor grading is a significant predictor of survival outcomes for this cohort.

From *Table 4*, it is possible to see that patients with well-differentiated tumors have the highest restricted mean survival time (14.68 months), indicating a favorable prognosis. Patients with intermediate-grade tumors have a moderate survival time (11.93 months), while poorly differentiated tumors have the lowest (10.43 months), suggesting the poorest prognosis.

The number of events is highest in the poorly differentiated group (22), followed by the intermediate group (19), and lowest in the well-differentiated group (7).

Confidence intervals show variability in survival times for intermediate and poorly differentiated groups, but upper bounds for all three categories and lower bound for well-differentiated one are unavailable, indicating that the survival curve does not provide enough information beyond the median survival time due to censoring or few events.

Up to now, we have provided some useful information on only the clinical covariates. In order to have a better understanding of the gene expression variables, in *Figure 2* we provide a correlation plot for said features.

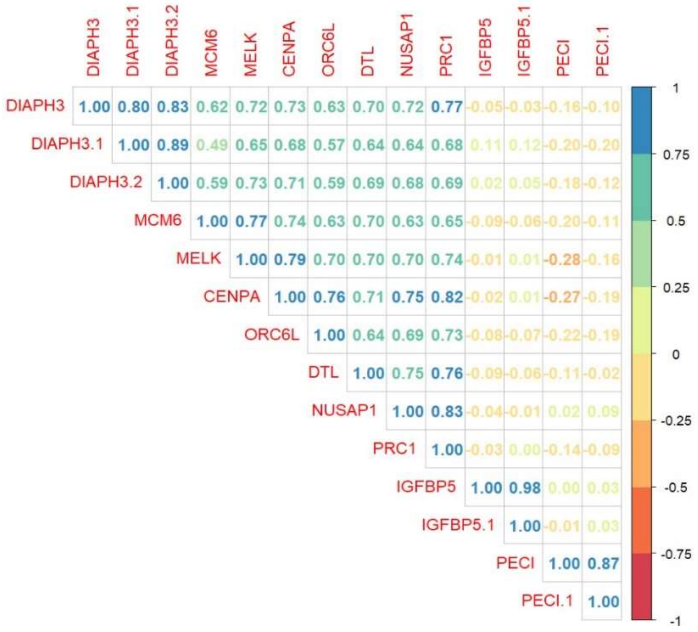

Figure 2

The above corrplot shows the relationships among various gene expression variables, indicating that genes **DIAPH3** and **DIAPH3.1** exhibit a strong correlation (0.80), indicating they are likely to be co-expressed. Similarly, **DIAPH3.1** and **DIAPH3.2** show an even higher correlation (0.89), suggesting a potentially regulatory relationship. **IGFBP5** and **IGFBP5.1** have an exceptionally high correlation (0.98), implying they may be nearly identical in their expression patterns or highly co-regulated.

As for moderate correlations, in *Figure 2* we can see that **MCM6** is moderately correlated with several genes: **CENPA** (0.74), **ORC6L** (0.63), and **DTL** (0.70).

On the other hand, **PECI** and **MELK** exhibit a negative correlation (-0.28), suggesting that their expression levels tend to vary inversely.

3. UNIVARIATE ANALYSIS on clinical variables

With the aim of inspecting the association between the clinical variables Diam, N, ER, Grade, Age and the outcome, i.e., the event of interest (metastasis or death, censoring) individually, the univariate analysis carried out using the Cox proportional hazards model can be found in *Table 5*.

HR	lower95%CI	upper95%CI	p	features
1.920	1.058	3.483	0.032	(Diam)>2cm
2.867	1.621	5.071	0.000	(N)>=4
0.485	0.256	0.919	0.027	(ER)Positive
1.486	0.462	0.802	0.194	(Grade)Poorly diff
2.753	1.100	0.208	0.081	(Grade)Well diff
0.942	0.896	0.991	0.020	Age

Table 5

As for **diameter**, the analysis results suggest that, given the hazard ratio of 1.920, patients with tumors larger than 2cm have a higher risk of metastasis or death compared to patients with tumors 2cm or smaller. The confidence interval indicates that we are 95% confident that the true hazard ratio lies between 1.058 and 3.483. The p-value of 0.032 suggests that the association between tumor diameter and risk of metastasis or death is statistically significant at a 0.05 level.

For patients with four or more **lymph nodes** involvement, the analysis shows that they have 2.867 times higher risk of metastasis or death compared to patients with fewer than four lymph nodes. The p-value smaller than 0.001 indicates a highly significant association between the number of positive lymph nodes and the risk of metastasis or death.

Patients positive for **estrogen receptor** status have a 0.485 times lower risk of death compared to those patients who resulted negative for estrogen receptor. The p-value of 0.027 suggests a statistically significant association between estrogen receptor status and the risk of metastasis or death.

As for tumor grading, the univariate analysis results do not indicate that there is a statistically significant association. However, these results might change when performing a multivariate analysis, given that the effects of said variables are considered simultaneously.

4. BASIC PREDICTIVE MODEL on clinical covariates

In order to gain insight into the factors potentially influencing patients' outcome, i.e., the event of interest, a Cox "basic" predictive model was developed, including the diameter of the tumor, the number of lymph nodes involved, the tumor's grading and the age of the patients.

```
## Call:
## coxph(formula = Surv(time, event) ~ factor(Diam) + factor(N) +
##     factor(ER) + factor(Grade) + Age, data = cancer)
##
##   n= 144, number of events= 48
##
##                             coef exp(coef) se(coef)      z Pr(>|z|)
## factor(Diam)>2cm          0.40347   1.49701  0.33061  1.220   0.2223
## factor(N)>=4              0.73700   2.08966  0.34171  2.157   0.0310 *
## factor(ER)Positive       -0.54480   0.57996  0.36723 -1.484   0.1379
## factor(Grade)Poorly diff  0.04725   1.04838  0.34374  0.137   0.8907
## factor(Grade)Well diff   -0.73628   0.47889  0.44919 -1.639   0.1012
## Age                      -0.04882   0.95235  0.02720 -1.795   0.0727 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                          exp(coef) exp(-coef) lower .95 upper .95
## factor(Diam)>2cm            1.4970     0.6680    0.7831     2.862
## factor(N)>=4                2.0897     0.4785    1.0696     4.083
## factor(ER)Positive          0.5800     1.7243    0.2824     1.191
## factor(Grade)Poorly diff    1.0484     0.9539    0.5345     2.056
## factor(Grade)Well diff      0.4789     2.0882    0.1986     1.155
## Age                         0.9523     1.0500    0.9029     1.005
##
## Concordance= 0.712  (se = 0.04 )
## Likelihood ratio test= 24.56  on 6 df,   p=4e-04
## Wald test            = 24.63  on 6 df,   p=4e-04
## Score (logrank) test = 26.79  on 6 df,   p=2e-04
```

Table 6

From the results of the model, it is possible to affirm that the tumor's diameter coefficient (0.403) suggests that patients with tumors larger than 2cm have a higher risk compared to patients with tumors 2cm or smaller.

Given that the coefficient equals 0.73, patients with four or more positive lymph nodes have about twice the risk compared to those with less than four positive lymph nodes.

The analysis also shows that estrogen receptor-positive patients have a lower risk compared to the estrogen receptor-negative ones, given that the coefficient equals 0.58.

As for the tumors grading, the results of the model show that there is no significant difference in their association with the risk of metastasis or death.

On the other hand, the coefficient for age shows that for each one-unit increase in patient's age, there is approximately a 0.95 times lower risk.

From the analysis it can also be stated that, among all predictors, only the number of lymph nodes involved, and the age of the patient are statistically significant in this model.

The p-value for all three tests, namely the Likelihood ratio test, the Wald test and the Score (logrank) test, suggest that, overall, the model is statistically significant in predicting the event.

5. PROPORTIONAL HAZARDS ASSUMPTION

In order to evaluate the functional form of the continuous variable, i.e., the age of the patients, we opted for the Martingale residuals plot and the Karnofski Performance Score vs. Hazard Ratio plot, to assess the linearity assumption.

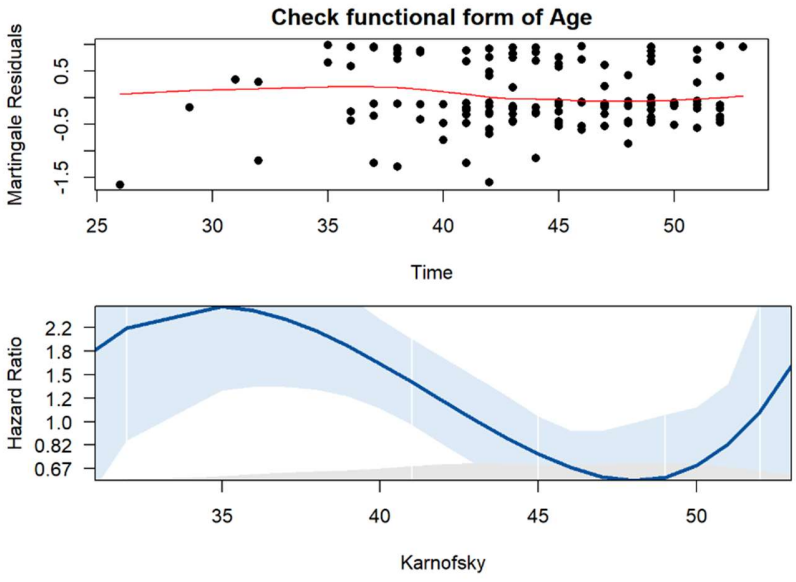

Figure 3

From the Martingale residual plot that can be seen in the above figure, we observe a slight presence of non-linearity. The residuals are not entirely randomly scattered around zero, and there are some

noticeable outliers. This suggests that the assumption of a linear relationship between age and the hazard may not be perfectly accurate.

This observation is further confirmed by the plot of Karnofsky Performance Score vs. Hazard Ratio, as it shows a clearly curved relationship, with the hazard ratio varying across different ages. Specifically, the hazard ratio starts high, decreases to its lowest point around age 47, and then increases again. The confidence interval, represented by the shaded area, is wider at the extremes, indicating greater uncertainty in the hazard ratio estimates for the youngest and oldest ages.

To access the proportional hazards assumption, we decided to check the Schoenfeld residuals plots of the diameter of the tumor, the number of lymph nodes involved, the tumor grading, the estrogen receptor status, and the age.

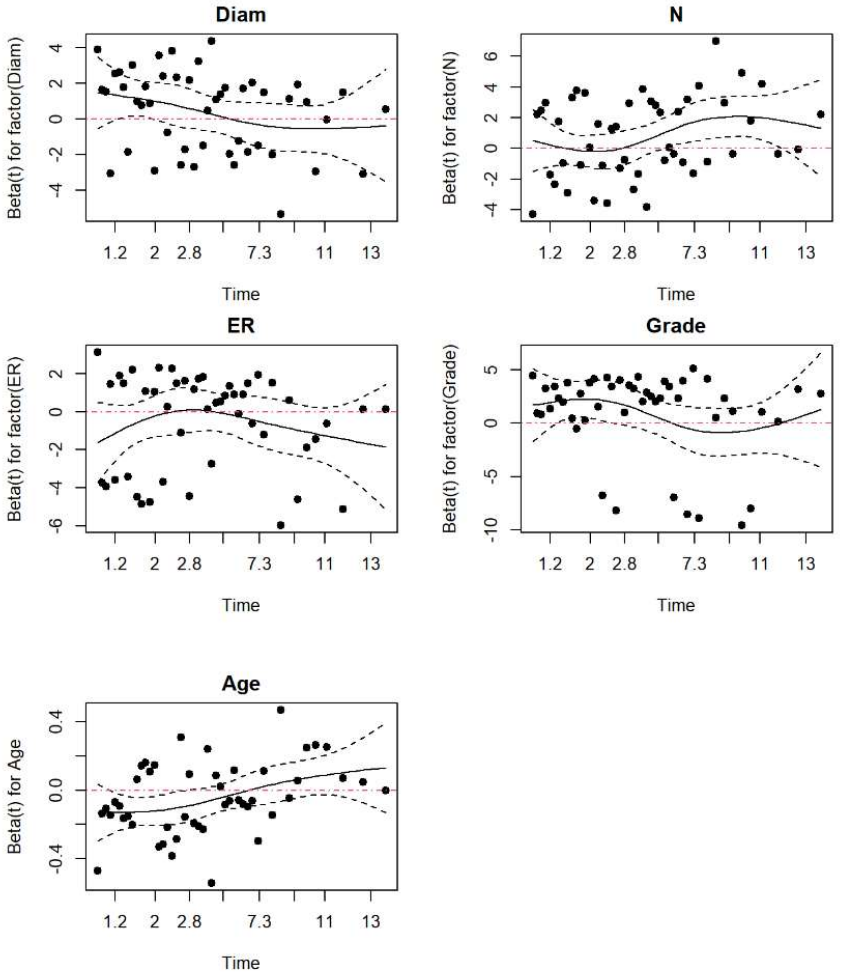

Figure 4

In *Figure 4*, we can observe that there is no strong evidence against the proportional hazards assumption for tumor diameter. The slight deviations are within the confidence bands, suggesting that the proportional hazards assumption holds for this covariate.

Analogously, there is no clear trend suggesting a violation of the assumption for the number of involved lymph nodes.

The assumption appears to be met for the estrogen receptor as well, as the residuals do not suggest any significant time-dependent effects.

While there are some variations, the overall pattern does not strongly suggest a violation of the assumption for tumor grading. The deviations are mostly within the confidence bands.

As it can be seen from the plot, there seems to be some indication that the proportional hazards assumption may be violated for age. This suggests that the effect of age on the hazard may not be constant over time, implying that the risk associated with age could change over the follow-up period.

To better assess the proportional hazards assumption for the categorical covariates, namely tumor diameter, number of lymph nodes involved, estrogen receptor status and tumor grading we are now going to take a look at the Log(-log(Survival) plots.

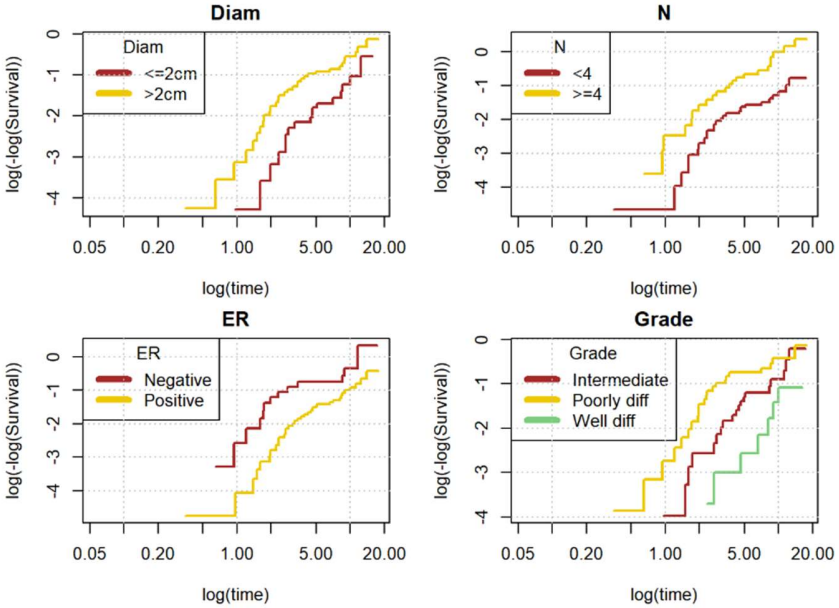

Figure 5

As for tumor diameter, the parallel nature of the lines in the plot suggest that the proportional hazards assumption is likely to be satisfied.

The assumption appears to be largely met for the number of involved lymph nodes as well, although some deviations are present.

There seems to be a slight divergence for ER, but overall, the parallel lines indicate a stable hazard ratio over time.

On the other hand, for tumor grading the assumption appears to not hold well, as the curves for patients having intermediated and poorly differentiated tumor appear to be intersecting at different points, indicating time-varying effects.

6. UNIVARIATE ANALYSIS ON GENE EXPRESSIONS & EVENT

To investigate the association of each gene expression variable with the outcome, i.e., the event of interest, the univariate analysis is shown below.

From the results of the analysis, we can see that statistically significant gene expressions are: Contig63649_RC, NUSAP1, QSCN6L1, SCUBE2, GMPS, ZNF533, RTN4RL1, Contig40831_RC, MELK, COL4A2, DTL, STK32B, ORC6L, RFC4, MS4A7, IGFBP5, IGFBP5.1, PRC1, CENPA and NM_004702.

By looking at the Hazard Ratios these significant gene expressions are associated with high risk of metastasis or death: PRC1, COL4A2, QSCN6L1, NUSAP1, RFC4, DTL, MELK, ORC6L, CENPA, NM_004702, Contig40831_RC, GMPS, Contig63649_RC, IGFBP5.1, IGFBP5 while these gene expressions are associated with low risk of metastasis or death: SCUBE2, ZNF533, MS4A7, RTN4RL1, STK32B.

	HR	lower95%CI	upper95%CI	p	features	significance
4	9.389	2.699	32.653	0.000	NUSAP1	**
7	9.574	3.005	30.501	0.000	QSCN6L1	**
64	16.283	4.736	55.986	0.000	PRC1	**
66	5.379	2.088	13.855	0.000	CENPA	**
33	0.255	0.116	0.562	0.001	ZNF533	**
49	6.429	2.084	19.834	0.001	ORC6L	**
68	4.920	1.765	13.712	0.002	NM_004702	**
60	3.149	1.483	6.685	0.003	IGFBP5.1	**
40	6.524	1.781	23.901	0.005	MELK	**
57	2.667	1.323	5.376	0.006	IGFBP5	**
50	7.492	1.690	33.212	0.008	RFC4	**
42	6.833	1.560	29.928	0.011	DTL	**
34	0.134	0.026	0.680	0.015	RTN4RL1	**
15	0.505	0.290	0.879	0.016	SCUBE2	**
41	11.381	1.535	84.397	0.017	COL4A2	**
43	0.077	0.009	0.646	0.018	STK32B	**
53	0.239	0.070	0.812	0.022	MS4A7	**
2	3.671	1.109	12.149	0.033	Contig63649_RC	**
38	4.637	1.121	19.178	0.034	Contig40831_RC	**
24	3.801	1.037	13.933	0.044	GMPS	**
51	2.003	0.991	4.049	0.053	CDCA7	
3	3.241	0.982	10.697	0.054	DIAPH3	
12	5.608	0.939	33.485	0.059	DIAPH3.2	
63	4.770	0.944	24.100	0.059	LGP2	
56	8.965	0.889	90.434	0.063	C9orf30	
54	3.715	0.821	16.806	0.088	MCM6	
23	3.107	0.814	11.857	0.097	ECT2	
36	0.255	0.050	1.301	0.100	PECI	
22	2.995	0.792	11.324	0.106	Contig35251_RC	
19	4.225	0.725	24.611	0.109	OXCT1	
69	2.476	0.808	7.592	0.113	ESM1	
18	2.668	0.773	9.211	0.121	GNAZ	
67	0.261	0.046	1.467	0.127	EGLN1	
9	2.525	0.764	8.344	0.129	DIAPH3.1	
20	2.205	0.785	6.195	0.133	MMP9	
14	3.382	0.655	17.460	0.146	C16orf61	
48	0.289	0.052	1.606	0.156	PECI.1	
45	3.765	0.593	23.915	0.160	FBXO31	
16	4.268	0.536	33.982	0.170	EXT1	
6	3.373	0.554	20.526	0.187	ALDH4A1	
37	2.929	0.542	15.825	0.212	MTDH	
39	0.509	0.176	1.472	0.213	TGFB3	
8	0.588	0.204	1.698	0.327	FGF18	
17	2.372	0.410	13.738	0.335	FLT1	
10	2.418	0.390	15.005	0.343	Contig32125_RC	
52	1.481	0.646	3.399	0.354	LOC643008	
27	0.476	0.092	2.458	0.375	CDC42RPA	
31	0.538	0.125	2.323	0.406	GPR180	
25	1.682	0.415	6.812	0.466	KNTC2	
32	1.395	0.540	3.601	0.492	RAB6B	
65	1.492	0.417	5.347	0.539	Contig20217_RC	
11	0.589	0.106	3.265	0.545	BBC3	
26	0.616	0.127	3.000	0.549	WISP1	
21	1.575	0.301	8.245	0.591	RUNDC1	
59	1.519	0.260	8.882	0.643	PITRM1	
5	1.494	0.262	8.527	0.652	AA555029_RC	
30	0.807	0.293	2.218	0.677	GSTM3	
13	0.754	0.198	2.868	0.678	RP5.860F19.3	
61	1.140	0.518	2.511	0.745	NMU	
29	1.437	0.161	12.850	0.746	AYTL2	
62	1.252	0.284	5.521	0.767	PALM2.AKAP2	
35	0.751	0.098	5.744	0.782	UCHL5	
1	1.112	0.465	2.660	0.811	TSPYL5	
55	0.809	0.143	4.591	0.811	AP2B1	
28	0.814	0.123	5.395	0.832	SERF1A	
58	1.065	0.308	3.684	0.921	HRASLS	
70	1.011	0.327	3.126	0.984	C20orf46	
44	0.989	0.146	6.691	0.991	DCK	
46	0.995	0.380	2.604	0.992	GPR126	
47	1.007	0.185	5.479	0.994	SLC2A3	

Table 7

7. VARIABLES SELECTION after Adjustment by Benjamini-Hochberg Method

To control for the false discovery rate when conducting multiple comparisons, in the following table we are going to have a look at the variables that remained statistically significant after adjusting by the Benjamini-Hochberg method.

	HR	lower95%CI	upper95%CI	p	features	q.BH	significance
33	0.255	0.116	0.562	0.001	ZNF533	0.0116667	**
57	2.667	1.323	5.376	0.006	IGFBP5	0.0420000	**
60	3.149	1.483	6.685	0.003	IGFBP5.1	0.0262500	**
68	4.920	1.765	13.712	0.002	NM_004702	0.0200000	**
66	5.379	2.088	13.855	0.000	CENPA	0.0000000	**
49	6.429	2.084	19.834	0.001	ORC6L	0.0116667	**
40	6.524	1.781	23.901	0.005	MELK	0.0388889	**
4	9.389	2.699	32.653	0.000	NUSAP1	0.0000000	**
7	9.574	3.005	30.501	0.000	QSCN6L1	0.0000000	**
64	16.283	4.736	55.986	0.000	PRC1	0.0000000	**

Table 8

From *Table 8*, we can observe that ten gene expressions remained statistically significant after the adjustment. Specifically, ZNF533 was associated with a lower risk of metastasis or death, with a hazard ratio (HR) of 0.255 and an adjusted p-value (q.BH) of ~ 0.011.

On the other hand, genes such as IGFBP5, IGFBP5.1, NM_004702, CENPA, ORC6L, MELK, NUSAP1, QSCN6L1, and PRC1 were associated with a higher risk of metastasis or death.

8. GENE EXPRESSION VARIABLES SELECTION WITH LASSO

In order to identify the most relevant genes associated with the outcome, i.e., the event of interest, we performed variable selection according to the Least Absolute Shrinkage and Selection Operator (Lasso) method.

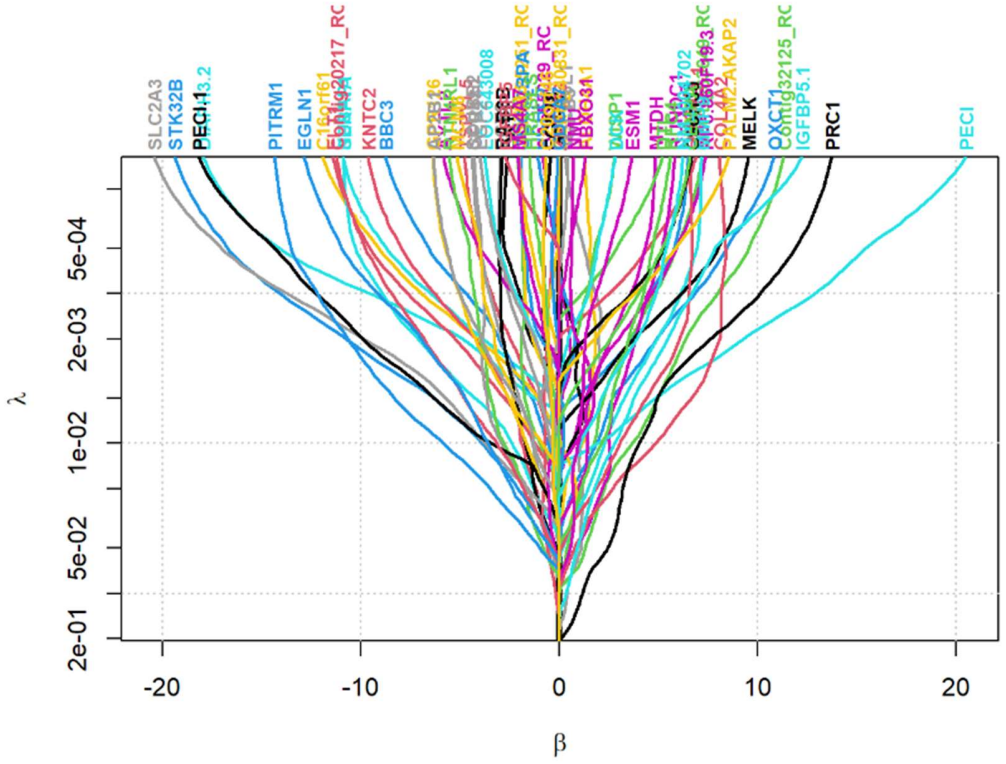

Figure 6

The LASSO regularization plot for gene expression variables reveals how gene coefficients change as the regularization parameter (λ) varies, helping identify significant predictors for the outcome. As λ increases, the penalty for including additional variables rises, causing less important gene coefficients to shrink towards zero. Genes like **STK32B**, **SLC2A3**, and **PECI** quickly approach zero, indicating they are less significant in predicting the outcome. In contrast, genes such as **PRC1**, **IGFBP5**, **IGFBP5.1**, **MELK**, **CENPA**, **NUSAP1**, and **QSCN6L1** maintain non-zero coefficients over a wide range of λ values, highlighting their strong predictive power. Specifically, **PRC1** and **IGFBP5** show robust coefficients, indicating substantial associations with the outcome.

To determine the optimal value of lambda (shrinkage parameter) that minimizes the cross-validation error, we are now going to look at the relationship between the partial likelihood deviance and the λ values.

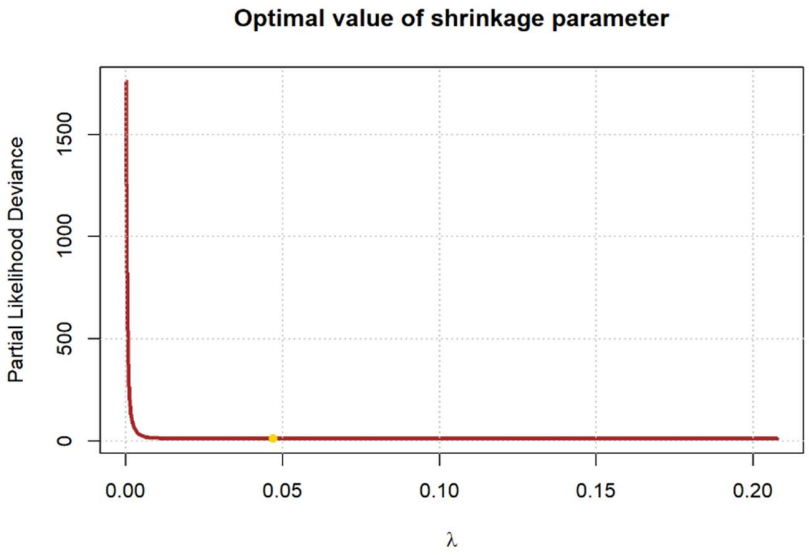

Figure 7

From the above plot, we can observe that the optimal λ appears to be slightly above 0.05, indicating that this minimizes the partial likelihood deviance, suggesting that this amount of regularization yields the most effective model.

The exact value of λ can be seen in *Figure 8*.

```
## [1] 0.04688611
```

Figure 8

From the variables selection by LASSO method, the gene expressions statistically significant in predicting the outcome can be seen in the following table:

```
##  [1] "Contig63649_RC" "NUSAP1"         "QSCN6L1"        "Contig32125_RC"
##  [5] "SCUBE2"         "OXCT1"          "MMP9"           "RUNDC1"
##  [9] "KNTC2"          "GPR180"         "RAB6B"          "ZNF533"
## [13] "RTN4RL1"        "Contig40831_RC" "COL4A2"         "STK32B"
## [17] "ORC6L"          "MS4A7"          "HRASLS"         "PITRM1"
## [21] "IGFBP5.1"       "PRC1"           "Contig20217_RC" "EGLN1"
## [25] "ESM1"
```

Table 9

9. AUGMENTED PREDICTIVE MODEL with clinical variables and gene expressions selected according to the LASSO method.

To gain insight into patients' outcomes, we built a predictive (Cox) model integrating clinical variables and statistically significant gene expression variables that were selected using the LASSO method. The results of the model can be seen in *Table 10*.

```
## Call:
## coxph(formula = Surv(time, event) ~ factor(Diam) + factor(N) +
##     factor(ER) + factor(Grade) + Age + Contig63649_RC + NUSAP1 +
##     QSCN6L1 + Contig32125_RC + SCUBE2 + OXCT1 + MMP9 + RUNDC1 +
##     KNTC2 + GPR180 + RAB6B + ZNF533 + RTN4RL1 + Contig40831_RC +
##     COL4A2 + STK32B + ORC6L + MS4A7 + HRASLS + PITRM1 + IGFBP5.1 +
##     PRC1 + Contig20217_RC + EGLN1 + ESM1, data = cancer)
##
##   n= 144, number of events= 48
##
##                               coef  exp(coef)   se(coef)      z Pr(>|z|)
## factor(Diam)>2cm          5.486e-02  1.056e+00  4.511e-01  0.122 0.903215
## factor(N)>=4              9.129e-02  1.096e+00  4.143e-01  0.220 0.825621
## factor(ER)Positive       -1.048e+00  3.506e-01  7.762e-01 -1.350 0.176879
## factor(Grade)Poorly diff  2.560e-01  1.292e+00  4.512e-01  0.567 0.570491
## factor(Grade)Well diff   -2.939e-02  9.710e-01  5.538e-01 -0.053 0.957671
## Age                      -7.476e-02  9.280e-01  4.227e-02 -1.769 0.076968 .
## Contig63649_RC            2.752e-02  1.028e+00  8.350e-01  0.033 0.973708
## NUSAP1                    2.116e+00  8.295e+00  1.656e+00  1.278 0.201368
## QSCN6L1                   1.766e+00  5.846e+00  1.331e+00  1.326 0.184755
## Contig32125_RC            3.949e+00  5.186e+01  1.359e+00  2.905 0.003669 **
## SCUBE2                   -1.002e+00  3.671e-01  7.078e-01 -1.416 0.156779
## OXCT1                    -8.744e-01  4.171e-01  1.666e+00 -0.525 0.599784
## MMP9                      2.110e+00  8.250e+00  9.705e-01  2.174 0.029676 *
## RUNDC1                    4.764e+00  1.172e+02  1.359e+00  3.505 0.000457 ***
## KNTC2                    -2.628e+00  7.219e-02  1.813e+00 -1.450 0.147099
## GPR180                   -1.806e+00  1.643e-01  1.479e+00 -1.221 0.222168
## RAB6B                    -1.896e+00  1.502e-01  8.627e-01 -2.197 0.027995 *
## ZNF533                   -1.098e+00  3.334e-01  5.699e-01 -1.927 0.053933 .
## RTN4RL1                  -1.921e+00  1.465e-01  1.383e+00 -1.389 0.164954
## Contig40831_RC            1.573e+00  4.822e+00  1.206e+00  1.304 0.192174
## COL4A2                    4.220e+00  6.801e+01  1.559e+00  2.707 0.006789 **
## STK32B                   -2.208e+00  1.099e-01  1.986e+00 -1.112 0.266175
## ORC6L                     3.062e+00  2.137e+01  1.406e+00  2.178 0.029431 *
## MS4A7                    -1.717e-02  9.830e-01  9.112e-01 -0.019 0.984963
## HRASLS                    1.154e+00  3.170e+00  1.057e+00  1.092 0.274800
## PITRM1                   -7.529e+00  5.374e-04  1.934e+00 -3.892 9.94e-05 ***
## IGFBP5.1                  2.165e+00  8.715e+00  6.938e-01  3.121 0.001805 **
## PRC1                      4.400e+00  8.148e+01  1.908e+00  2.307 0.021064 *
## Contig20217_RC           -2.548e+00  7.820e-02  1.408e+00 -1.810 0.070333 .
## EGLN1                    -4.395e+00  1.234e-02  1.588e+00 -2.768 0.005632 **
## ESM1                      1.187e+00  3.278e+00  9.991e-01  1.188 0.234699
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                          exp(coef) exp(-coef) lower .95 upper .95
## factor(Diam)>2cm         1.056e+00  9.466e-01 4.363e-01 2.558e+00
## factor(N)>=4             1.096e+00  9.128e-01 4.864e-01 2.468e+00
## factor(ER)Positive       3.506e-01  2.852e+00 7.657e-02 1.605e+00
## factor(Grade)Poorly diff 1.292e+00  7.741e-01 5.334e-01 3.128e+00
## factor(Grade)Well diff   9.710e-01  1.030e+00 3.280e-01 2.875e+00
## Age                      9.280e-01  1.078e+00 8.542e-01 1.008e+00
## Contig63649_RC           1.028e+00  9.729e-01 2.001e-01 5.280e+00
## NUSAP1                   8.295e+00  1.206e-01 3.231e-01 2.129e+02
## QSCN6L1                  5.846e+00  1.711e-01 4.301e-01 7.946e+01
## Contig32125_RC           5.186e+01  1.928e-02 3.614e+00 7.442e+02
## SCUBE2                   3.671e-01  2.724e+00 9.169e-02 1.470e+00
## OXCT1                    4.171e-01  2.397e+00 1.591e-02 1.093e+01
## MMP9                     8.250e+00  1.212e-01 1.231e+00 5.528e+01
## RUNDC1                   1.172e+02  8.530e-03 8.167e+00 1.683e+03
## KNTC2                    7.219e-02  1.385e+01 2.067e-03 2.521e+00
## GPR180                   1.643e-01  6.086e+00 9.047e-03 2.985e+00
## RAB6B                    1.502e-01  6.657e+00 2.770e-02 8.148e-01
## ZNF533                   3.334e-01  2.999e+00 1.091e-01 1.019e+00
## RTN4RL1                  1.465e-01  6.826e+00 9.737e-03 2.204e+00
## Contig40831_RC           4.822e+00  2.074e-01 4.534e-01 5.129e+01
## COL4A2                   6.801e+01  1.470e-02 3.204e+00 1.444e+03
## STK32B                   1.099e-01  9.099e+00 2.242e-03 5.388e+00
## ORC6L                    2.137e+01  4.680e-02 1.358e+00 3.362e+02
## MS4A7                    9.830e-01  1.017e+00 1.648e-01 5.863e+00
## HRASLS                   3.170e+00  3.154e-01 3.997e-01 2.515e+01
## PITRM1                   5.374e-04  1.861e+03 1.213e-05 2.382e-02
## IGFBP5.1                 8.715e+00  1.147e-01 2.237e+00 3.395e+01
## PRC1                     8.148e+01  1.227e-02 1.938e+00 3.425e+03
## Contig20217_RC           7.820e-02  1.279e+01 4.950e-03 1.236e+00
## EGLN1                    1.234e-02  8.105e+01 5.495e-04 2.771e-01
## ESM1                     3.278e+00  3.051e-01 4.626e-01 2.323e+01
##
## Concordance= 0.906  (se = 0.018 )
## Likelihood ratio test= 128.7  on 31 df,   p=7e-14
## Wald test            = 76.76  on 31 df,   p=9e-06
## Score (logrank) test = 138.2  on 31 df,   p=2e-15
```

Table 10

From the results of the model, we can observe that clinical features such as tumor diameter, lymph node count, ER status, and tumor grade didn't show strong associations with outcomes, with high p-values indicating their limited predictive value. However, age showed a marginal significance. In contrast, certain genes exhibited significant associations, indicating the importance of molecular data in predictive modeling.

The augmented Cox model, incorporating both clinical and gene expression variables, showed strong predictive ability with a high Concordance Index (C-index) and significant test results. Genes such as **Contig63649_RC, COL4A2, ORC6L, PRC1**, and **PITRM1** emerged as significant predictors.

Among the significant gene expressions, several are associated with a high risk of metastasis or death, including **Contig32125_RC, MMP9, RUNDC1, COL4A2, ORC6L**, and **IGFBP5.1**. Conversely, expressions such as **RAB6B, PITRM1**, and **ESM1** are linked to a lower risk.

The p-value for all three tests, namely the Likelihood ratio test, the Wald test and the Score (logrank) test, suggest that, overall, the model is statistically significant in predicting the event.

10. EVENT RISK PREDICTION

The accurate prediction of an event risk at a fixed time-point is crucial for effective clinical decision-making and personalized treatment planning for patients.

For this reason, we are now going to observe the results obtained from the predictive analysis carried out to predict the risk of the event (metastasis or death) at a fixed time-point, i.e., 12 months of follow-up, for three patients that were randomly selected from the cohort.

The analysis was carried out using two predictive models; specifically, in the following we can observe the results obtained from a “basic” predictive (Cox) model, which relies only on the clinical covariates (diameter of the tumor, number of lymph nodes involved, tumor grading and age), and the results obtained from the augmented model, which includes, along with the clinical variables, also gene expression variables previously selected according to their statistical significance.

Therefore, *Table 10*, we can take a look at the results.

risk_base_model	risk_agu_model	Diam	N	ER	Grade	Age	Contig63649_RC	NUSAP1	QSCN6L1	Contig32125_RC	SCUBE2	OXCT1	MMP9	RUNDC1	KNTC2	GPR180	RAB6B
0.5985822	0.0958204	>2cm	<4	Negative	Intermediate	44	0.1748558	-0.0665893	0.0002217	-0.1290805	-0.7396632	0.0534161	-0.2053575	-0.2020483	0.3059753	0.0701824	0.1857224
0.1723223	0.3941060	>2cm	<4	Positive	Well diff	50	-0.2483648	0.0673382	-0.0762002	-0.3035071	-0.5304818	-0.1133317	-0.3403279	-0.1660510	-0.0580757	0.0285566	0.0455405
0.3489703	0.5247076	<=2cm	<4	Positive	Poorly diff	41	0.0941969	0.4577536	-0.0834792	-0.1716594	-0.6876388	0.1058986	-0.3180016	-0.1613048	0.0541958	0.1158649	-0.0806325

ZNF533	RTN4RL1	Contig40831_RC	COL4A2	STK32B	ORC6L	MS4A7	HRASLS	PITRM1	IGFBP5.1	PRC1	Contig20217_RC	EGLN1	ESM1
-0.5307709	-0.0257542	0.1275547	0.0665459	0.1382768	0.2831944	-0.2743678	-0.2965329	-0.0091776	-0.1072163	-0.0485553	0.2765233	0.0710948	-0.4467877
0.0487754	-0.0121821	-0.1199183	0.0774409	-0.1232907	0.1805166	0.0857294	-0.0400482	-0.0355186	-0.1258986	0.1697266	-0.2750999	0.0414185	-0.1434093
-0.5628388	0.3183525	-0.2566424	-0.0742114	-0.0418261	0.2943387	-0.2999745	-0.0091527	-0.1495010	-0.1946291	0.1838662	0.2489721	0.2338470	0.1460169

Table 10

As it can be seen from the results of the comparison between the two models, the base model predicts a risk of 0.5896 for the **first patient**, which drops significantly to 0.0958 with the augmented model. It is important to note that the first patient, aged 44, has a tumor diameter greater than 2 cm, is ER-negative, and has an intermediate grade tumor. The drastic reduction in risk from the base model to the augmented model suggests that additional features used in the augmented model, i.e., selected gene expression variables, may have identified mitigating factors not considered in the base model.

As for the **second patient**, the base model predicts a risk of 0.1723, which increases to 0.3941 with the augmented model. This patient, aged 50, has a tumor diameter greater than 2 cm, is ER-positive, and has a well-differentiated tumor. The increase in risk in the augmented model suggests that the additional genetic features likely contribute to a higher risk assessment despite the favorable ER status and well-differentiated grade.

The base model for the **third patient** indicates a risk of 0.3490, which increases to 0.5247 with the augmented model. The third patient, aged 41, has a tumor diameter of 2 cm or less, is ER-positive, and has a poorly differentiated tumor. The higher risk predicted by the augmented model reflects the impact of poor differentiation and possibly additional genetic factors that elevate the overall risk.

Therefore, from the observed results it is possible to affirm that the augmented model's predictions vary significantly from the base model, highlighting the importance of integrating additional clinical and genetic data to refine risk assessments. The differences between the two models suggest that the augmented model provides a more nuanced understanding of individual risk factors, potentially leading to better-informed clinical decisions.

11. CONCLUSIONS

In this project we aimed to analyze a dataset derived from an observational study focused on metastasis-free survival in 144 women diagnosed with breast cancer who have lymph node involvement, to derive insights and predictive models to enhance our understanding and management of breast cancer progression.

The dataset includes a range of clinical and gene expression variables, offering a comprehensive view of factors that may influence patient outcomes in terms of their response to treatment.

The descriptive analysis revealed that patients' age at diagnosis ranged from 26 to 53 years, with a median age of 45 years. Tumor size was evenly distributed, with 73 women having tumors ≤ 2 cm and 71 having tumors > 2 cm. Most women (106) had fewer than 4 involved lymph nodes, indicating a relatively early stage of lymph node involvement for most patients. The cohort was predominantly estrogen receptor-positive (117 women), which is associated with better responsiveness to hormonal therapies and a favorable prognosis. Tumor grade distribution varied, with 38% intermediate, 33% poorly differentiated, and 28% well-differentiated, highlighting cohort heterogeneity and varying prognostic implications. Follow-up times varied from 0.05 to 18 months, with a median of 7 months. Among 144 patients, 96 were censored, while 48 experienced metastasis or death, indicating a relatively favorable short-term prognosis for the majority.

Further analysis of follow-up outcomes showed that during the first half of the follow-up period (8.8 months), 55 women were censored, while 41 experienced metastasis or death. By the end of the follow-up period, 41 women were censored, and 7 experienced metastasis or death. The analysis of tumor grading indicated that patients with intermediate-grade tumors had a balanced distribution in tumor size and predominantly positive estrogen receptor status. Poorly differentiated tumors showed larger sizes and a higher proportion of estrogen receptor-negative cases, with a higher rate of adverse events (46% experiencing metastasis or death). Well-differentiated tumors had the smallest sizes, predominantly positive estrogen receptor status, fewer involved lymph nodes, and the lowest rate of adverse events (17%).

The analysis of the Kaplan-Meier survival curves indicated that patients with well-differentiated tumors had the highest survival probability, while those with poorly differentiated tumors had the lowest. The differences in survival probabilities among the three tumor grades were statistically significant ($p = 0.019$), suggesting that tumor grading is a significant predictor of survival

outcomes. Restricted mean survival times also varied, with well-differentiated tumors showing the highest survival time (14.68 months), followed by intermediate-grade tumors (11.93 months), and poorly differentiated tumors (10.43 months).

The univariate Cox proportional hazards analysis indicated that larger tumor diameter, greater lymph node involvement, and estrogen receptor-negative status were associated with higher risk of metastasis or death.

To assess the proportional hazards assumption, Schoenfeld residuals plots and Log(-log(Survival)) plots were examined, indicating that the assumption largely held for most covariates except age.

The univariate analysis of gene expression variables identified several genes significantly associated with the outcome, and after adjustment for multiple testing using the Benjamini-Hochberg method, ten genes remained statistically significant.

Using the Least Absolute Shrinkage and Selection Operator (LASSO) method for variable selection, genes such as PRC1, IGFBP5, and others were identified as significant predictors. An augmented predictive model, integrating both clinical and significant gene expression variables, showed strong predictive ability with a high concordance index (C-index). The model highlighted genes like Contig63649_RC, COL4A2, and ORC6L as significant predictors of metastasis or death.

Finally, from the comparison of risk predictions of basic and augmented models for three randomly selected patients, results showed that the augmented model provided a more nuanced understanding of individual risk factors. The results underscored the importance of integrating genetic data with clinical variables to refine risk assessments, potentially leading to better-informed clinical decisions and personalized treatment planning for patients.