KU LEUVEN

# Statistical Analysis of Reliability and Survival Data: Rotterdam Dataset

**Authors**
Aksoy, Barış r0869901
Heller, Jack r0862809
Zdravković, Aleksandra r0869484

Leuven. May, 2022

# 1 Exploratory Data Analysis

These data sets are used in the paper by Royston and Altman that is referenced below. The Rotterdam data is used to create a fitted model, and the GBSG data for validation of the model. The paper gives references for the data source.

There are 43 subjects who have died without recurrence, but whose death time is greater than the censoring time for recurrence. A common way that this happens is that a death date is updated in the health record sometime after the research study ended, and said value is then picked up when a study data set is created. But it raises serious questions about censoring. For instance subject 40 is censored for recurrence at 4.2 years and died at 6.6 years; when creating the endpoint of recurrence free survival (earlier of recurrence or death), treating them as a death at 6.6 years implicitly assumes that they were recurrence free just before death. For this to be true we would have to assume that if they had progressed in the 2.4 year interval before death (while off study), that this information would also have been noted in their general medical record, and would also be captured in the study data set. However, that may be unlikely. Death information is often in a centralized location in electronic health records, easily accessed by a programmer and merged with the study data, while recurrence may require manual review. How best to address this is an open issue.

Table 1: Data description

| | |
|---|---|
| pid | Patient identifier |
| year | Year of surgery |
| age | Age at surgery |
| meno | Menopausal status (0 = premenopausal, 1 = postmenopausal) |
| size | Tumor size, a factor with levels <= 20, 20-25, > 50 |
| grade | Differentiation grade |
| nodes | Number of positive lymph nodes |
| pgr | Progesterone receptors (fmol/l) |
| er | Estrogen receptors (fmol/l) |
| hormon | Hormonal treatment (0=no, 1=yes) |
| chemo | Chemotherapy |
| rtime | Days to relapse or last follow-up |
| recur | 0 = no relapse, 1 = relapse |
| dtime | Days to death or last follow-up |
| death | 0 = alive, 1 = dead |

Table 1 explains the covariates in the Rotterdam dataset.

# 2 Further Analysis

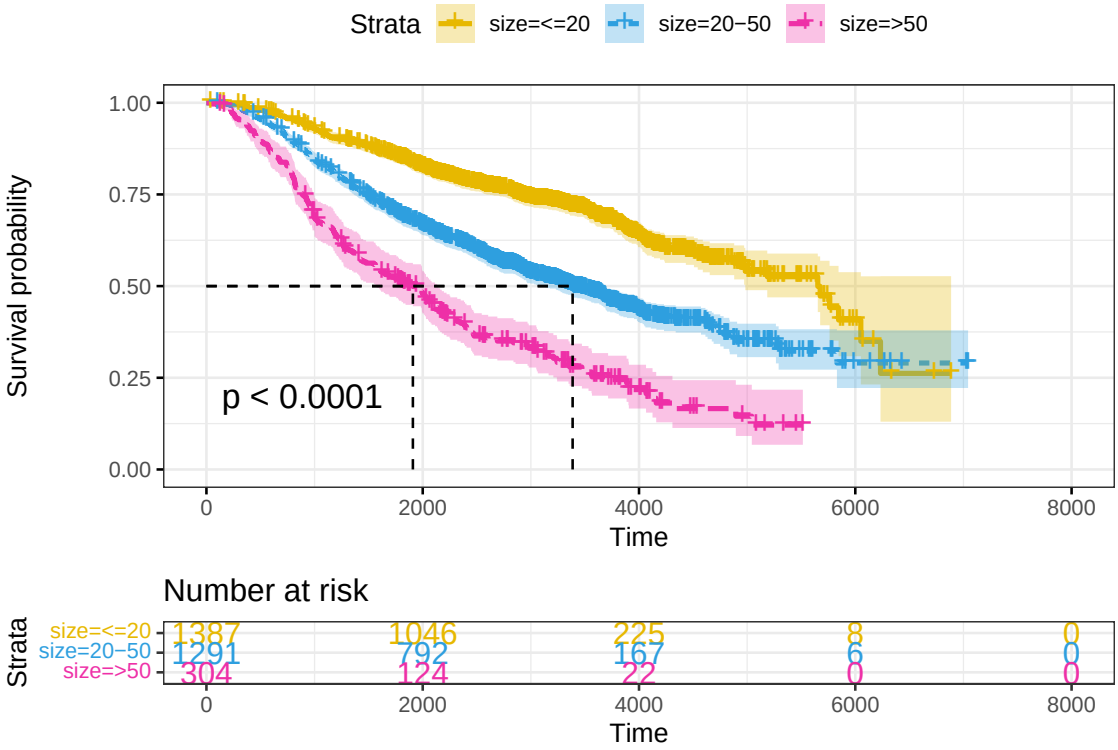
Now the focus will be on the response variable, the censoring indicator, and the categorical variable.

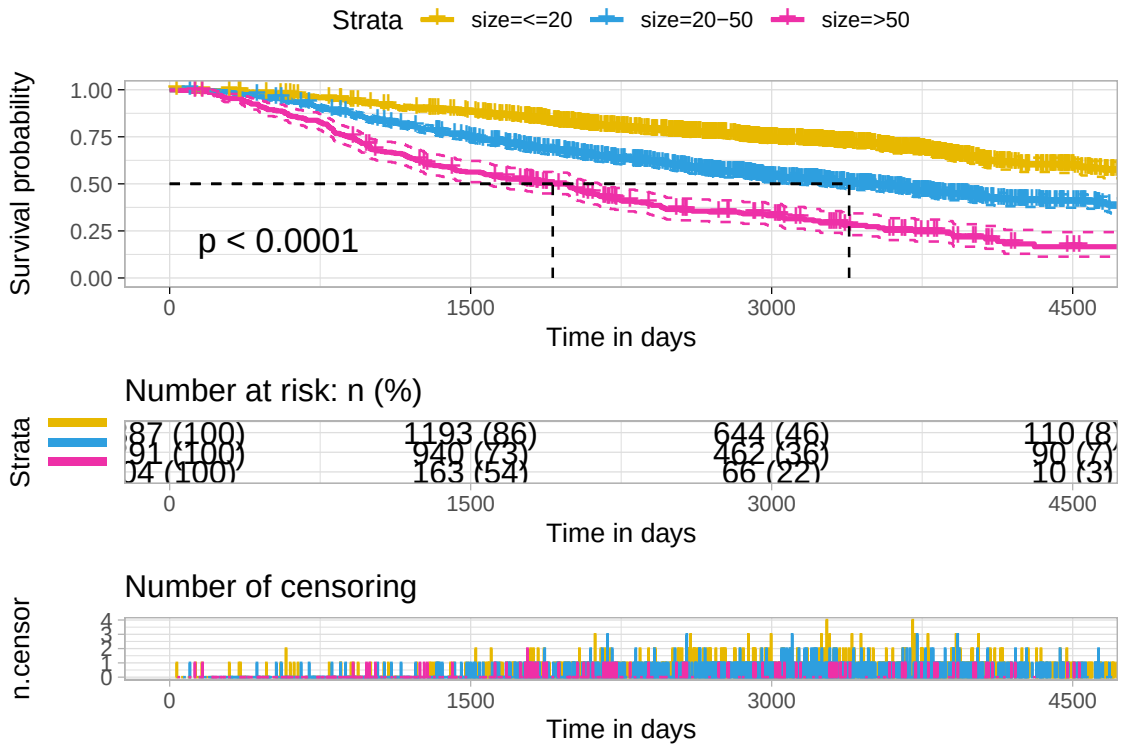
## 2.1 Survival Distrubution by Levels of ...

(a) For each of the levels of the categorical variable, compute the survival distribution. Plot them on the same graph. What do the graphs suggest?

### 2.1.1 Size

```
##   time n.risk n.event n.censor      surv     upper     lower
## 1   36   1387       0        1 1.0000000 1.0000000 1.0000000
## 2   64   1386       1        0 0.9992785 1.0000000 0.9978659
## 3   97   1385       1        0 0.9985570 1.0000000 0.9965606
## 4  101   1384       1        0 0.9978355 1.0000000 0.9953918
## 5  129   1383       0        1 0.9978355 1.0000000 0.9953918
## 6  141   1382       1        0 0.9971135 0.9999421 0.9942928
```






The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

Table 2: Summary of the model.

| | records | n.max | n.start | events | rmean | se(rmean) | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|---|---|---|
| Size (≤20) | 1387 | 1387 | 1387 | 414 | 4721.199 | 119.40159 | 5653 | 4983 | |
| Size (20−50) | 1291 | 1291 | 1291 | 646 | 3807.025 | 95.52799 | 3386 | 3084 | 3690 |
| Size (>50) | 304 | 304 | 304 | 212 | 2537.178 | 148.10071 | 1909 | 1566 | 2141 |

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2250, the probability of survival is approximately 0.625 for size>=50, and 0.85 for size<50.

From Table 2 can be seen that the median survival for Size  50 is 5653, for Size 20-50 is 3386, and for Size > 50 is 1909. This suggests slightly worse survival for patients with tumor of larger size. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

### 2.1.2 Menopause

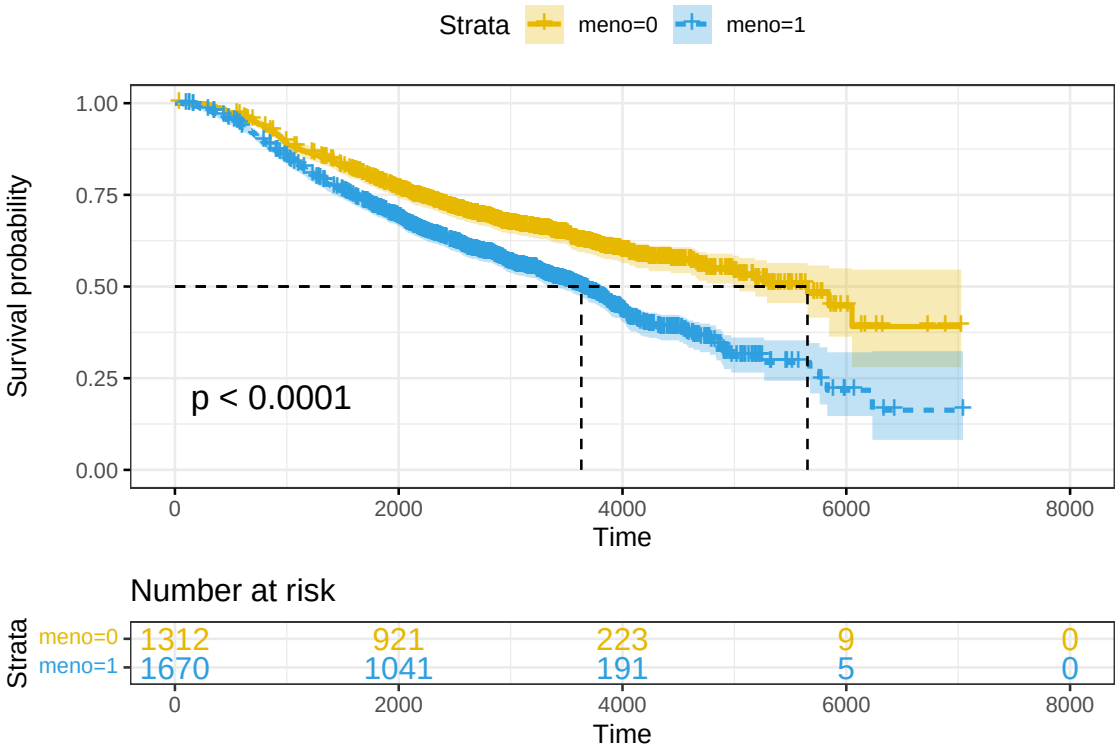



Table 3: Median survival times for each group.

| | records | n.max | n.start | events | rmean | se(rmean) | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|---|---|---|
| Meno = 0 | 1312 | 1312 | 1312 | 468 | 4622.427 | 108.80722 | 5653 | 4983 | |
| Meno = 1 | 1670 | 1670 | 1670 | 804 | 3672.887 | 92.46545 | 3632 | 3368 | 3813 |

The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for premenopausal patients, and 0.825 for postmenopausal patients.

From Table 3 can be seen that the median survival is 5653 for premenopausal patients, and 3632 for postmenopausal, suggesting slightly worse survival for patients that have gone through menopause. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

### 2.1.3 Hormonal Treatment

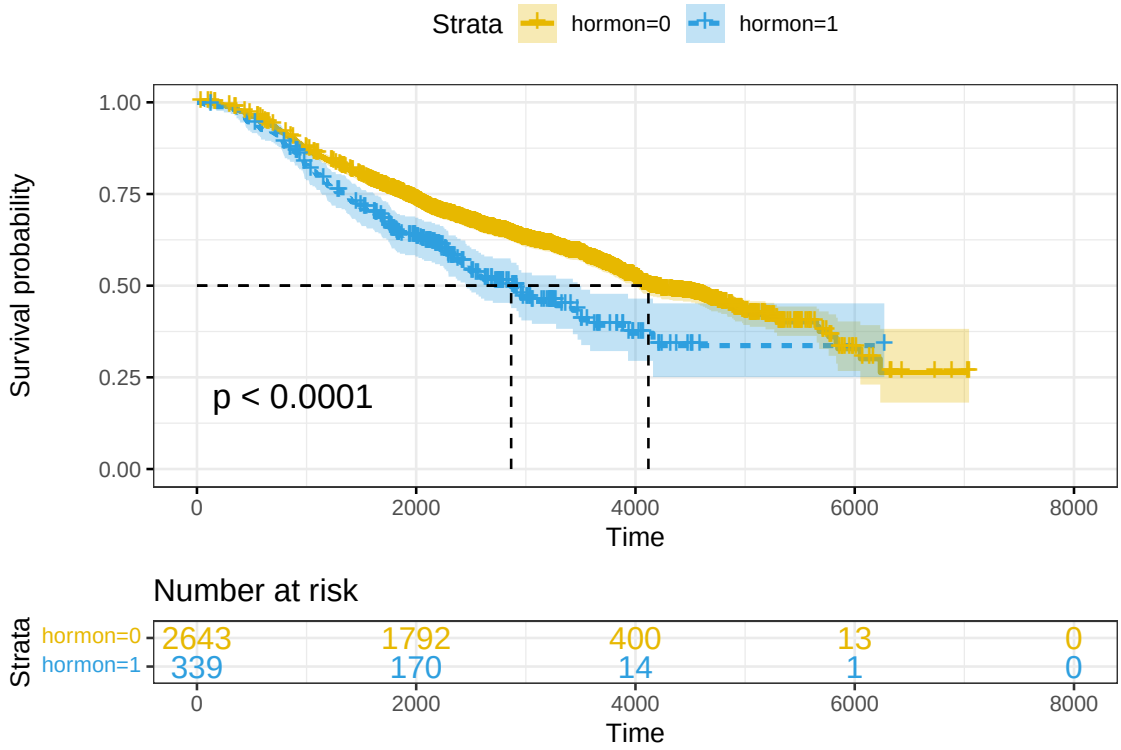



Table 4: Median survival times for each group.

| | records | n.max | n.start | events | rmean | se(rmean) | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|---|---|---|
| hormon=0 | 2643 | 2643 | 2643 | 1113 | 4159.588 | 76.57099 | 4118 | 3988 | 4614 |
| hormon=1 | 339 | 339 | 339 | 159 | 3659.665 | 203.38456 | 2866 | 2450 | 3472 |

The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for premenopausal patients, and 0.825 for postmenopausal patients.

From Table 4 can be seen that the median survival is 4118 for patients that went through hormonal therapy, and 2866 for those who did not, suggesting slightly worse survival for patients that have gone through the hormonal therapy. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

### 2.1.4 Chemotherapy

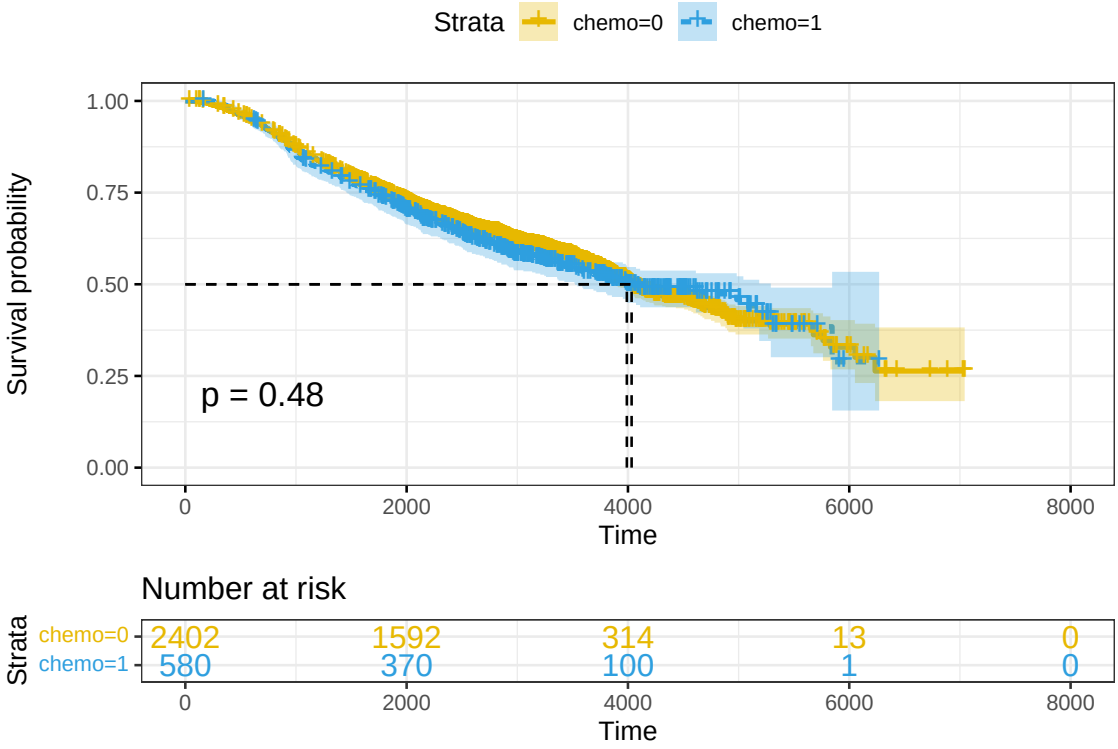



Table 5: Median survival times for each group.

| | records | n.max | n.start | events | rmean | se(rmean) | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|---|---|---|
| chemo=0 | 2402 | 2402 | 2402 | 1014 | 4103.663 | 80.06765 | 4033 | 3885 | 4239 |
| chemo=1 | 580 | 580 | 580 | 258 | 4080.311 | 162.59049 | 3990 | 3522 | 5291 |

The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for premenopausal patients, and 0.825 for postmenopausal patients.

From Table 5 can be seen that the median survival is 4118 for patients that went through hormonal therapy, and 2866 for those who did not, suggesting slightly better survival for patients that have gone through chemotherapy. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

## 2.2 Confidence Intervals and Estimators by Levels of ...

(b) For each level obtain an appropriate estimator and confidence interval for the 3 quartiles of the survival curves. Interpret the results.

## 2.3 Test of Differences Between the Survival Curves

(c) Conduct a single test of differences between the survival curves. Justify your choice of test.

Table 6: Estimates and confidence intervals for 3 quartiles for each level of covariates *size, chemo, meno, hormon.*

| | $\widehat{q_1}$ | 5% | 95% | $\widehat{q_2}$ | 5% | 95% | $\widehat{q_3}$ | 5% | 95% |
|---|---|---|---|---|---|---|---|---|---|
| Size ($\leq$50) | 2880 | 2590 | 3315 | 5653 | 4983 | *NA* | *NA* | 6051 | *NA* |
| Size (20−50) | 1476 | 1361 | 1623 | 3386 | 3084 | 3690 | *NA* | 5830 | *NA* |
| Size (>50) | 890 | 809 | 999 | 1909 | 1566 | 2141 | 3714 | 3240 | 4309 |
| Chemo = 0 | 1812 | 1677 | 1957 | 4033 | 3885 | 4239 | *NA* | 6051 | *NA* |
| Chemo = 1 | 1699 | 1455 | 1954 | 3990 | 3522 | 5291 | *NA* | 5845 | *NA* |
| Meno = 0 | 2115 | 1944 | 2371 | 5653 | 4983 | *NA* | *NA* | *NA* | *NA* |
| Meno = 1 | 1571 | 1424 | 1723 | 3632 | 3368 | 3813 | 5762 | 5266 | *NA* |
| Hormon = 0 | 1882 | 1742 | 1994 | 4118 | 3988 | 4614 | *NA* | 6051 | *NA* |
| Hormon = 1 | 1361 | 1140 | 1618 | 2866 | 2450 | 3472 | *NA* | *NA* | *NA* |

### 2.3.1 I need to rephrase all of this text, because it's been copy-pasted!!!

Now, the questions that arises is if these two curves are statistically equivalent. For answering it, we can use the log-rank test (Mantel 1966; Peto and Peto 1972). This is the most well-known and widely used method to test the null hypothesis of no difference in survival between two or more independent groups. It is a large-sample chi-square test that is obtained by constructing a two by two contingency table at each distinct event time, and comparing the failure rates between the two groups, conditional on the number at risk in each group. The test compares the entire survival experience between groups and can be thought of as a test of whether the survival curves are identical or not.

When we state that two KM curves are statistically equivalent, we mean that, based on a testing procedure that compares the two curves in some overall sense, we do not have evidence to indicate that the true (population) survival curves are different. The null hypothesis of the testing procedure is that there is no overall difference between the two (or k) survival curves. Under this, the log–rank statistic is approximately a chi-square with k−1 degree of freedom. Thus, tables of the chi-square distribution are used to determine the p-value. This test is the one with most power to test differences that fit the proportional hazards model - so works well as a set-up for subsequent Cox regression. It gives equal weight to early and late failures.

An alternative test that is often used is the Peto & Peto (Peto and Peto 1972) modification of the Gehan-Wilcoxon test (Gehan 1965). This last one is a variation of the log-rank test statistic and is derived by applying different weights at the f−th failure time. This approach is most sensitive to early differences (or earlier time points) between survival. This type of weighting may be used to assess whether the effect of a treatment/marketing campaign on survival is strongest in the earlier phases of administration/contacto and tends to be less effective over time. (Marta Sestelo 2017)

Table 7: Stratified log-rank test for differences in menopause.

| | $\chi^2$ | df | p-value |
|---|---|---|---|
| Stratified by Age | 12.05 | 1 | 0.0005177 |
| Stratified by Size | 42.53 | 1 | 0.0000000 |
| Stratified by Hormon | 44.06 | 1 | 0.0000000 |

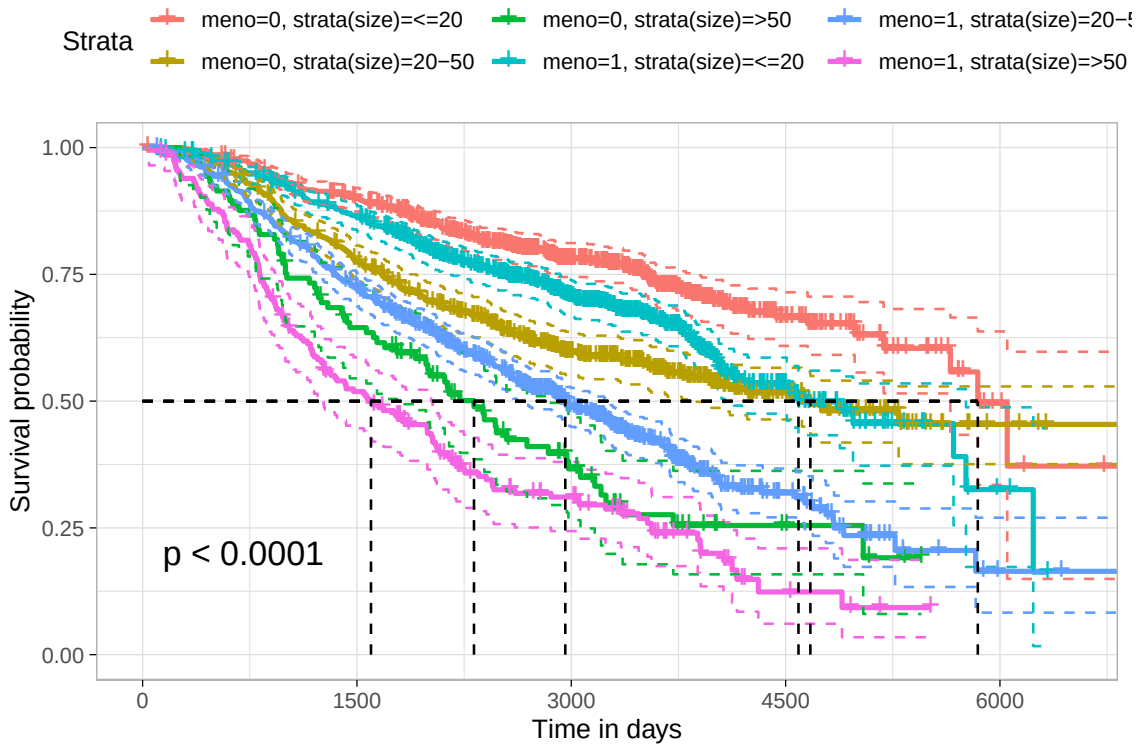



Figure 1: Independent variable meno stratified by size of the tumor.

# 3 Modeling

## 3.1 Semiparametric PH modeling

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Call:
## coxph(formula = Surv(dtime, as.numeric(death)) ~ chemo + meno +
##     hormon + age + as.factor(grade) + nodes + pgr, data = data)
##
##   n= 2982, number of events= 1272
##
##                          coef  exp(coef)   se(coef)      z Pr(>|z|)
## chemo                0.1009031  1.1061695  0.0809475  1.247 0.212572
## meno                -0.0215349  0.9786954  0.0992744 -0.217 0.828269
## hormon              -0.0320038  0.9685029  0.0883273 -0.362 0.717104
## age                  0.0189625  1.0191434  0.0038079  4.980 6.37e-07 ***
## as.factor(grade)3    0.3751328  1.4551847  0.0705124  5.320 1.04e-07 ***
## nodes                0.0869173  1.0908065  0.0044943 19.339  < 2e-16 ***
```

```
## pgr               -0.0004244  0.9995757  0.0001200 -3.536 0.000406 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## chemo                1.1062     0.9040    0.9439    1.2964
## meno                 0.9787     1.0218    0.8056    1.1889
## hormon               0.9685     1.0325    0.8145    1.1516
## age                  1.0191     0.9812    1.0116    1.0268
## as.factor(grade)3    1.4552     0.6872    1.2674    1.6709
## nodes                1.0908     0.9168    1.0812    1.1005
## pgr                  0.9996     1.0004    0.9993    0.9998
##
## Concordance= 0.681  (se = 0.008 )
## Likelihood ratio test= 437.5  on 7 df,   p=<2e-16
## Wald test            = 552.3  on 7 df,   p=<2e-16
## Score (logrank) test = 606.8  on 7 df,   p=<2e-16

##                            chemo          meno        hormon           age
## chemo               6.552502e-03  1.369306e-03  3.678275e-04  5.210349e-05
## meno                1.369306e-03  9.855399e-03 -7.245988e-04 -2.798504e-04
## hormon              3.678275e-04 -7.245988e-04  7.801705e-03 -6.887004e-06
## age                 5.210349e-05 -2.798504e-04 -6.887004e-06  1.450015e-05
## as.factor(grade)3 -9.005083e-05 -3.364951e-04 -2.273317e-04  6.833926e-06
## nodes             -7.696332e-05 -4.263381e-05 -5.459167e-05  6.158417e-08
## pgr               -3.604232e-07  1.062970e-06  3.914411e-07 -4.541636e-08
##                   as.factor(grade)3         nodes           pgr
## chemo                -9.005083e-05 -7.696332e-05 -3.604232e-07
## meno                 -3.364951e-04 -4.263381e-05  1.062970e-06
## hormon               -2.273317e-04 -5.459167e-05  3.914411e-07
## age                   6.833926e-06  6.158417e-08 -4.541636e-08
## as.factor(grade)3     4.972000e-03 -1.737616e-05  1.129103e-06
## nodes                -1.737616e-05  2.019870e-05  2.341796e-08
## pgr                   1.129103e-06  2.341796e-08  1.440583e-08

## Call:
## coxph(formula = Surv(dtime, as.numeric(death)) ~ chemo + meno +
##     hormon + age + as.factor(grade) + nodes + pgr, data = data)
##
##   n= 2982, number of events= 1272
##
##                        coef  exp(coef)   se(coef)      z Pr(>|z|)
## chemo              0.1009031  1.1061695  0.0809475  1.247 0.212572
## meno              -0.0215349  0.9786954  0.0992744 -0.217 0.828269
## hormon            -0.0320038  0.9685029  0.0883273 -0.362 0.717104
## age                0.0189625  1.0191434  0.0038079  4.980 6.37e-07 ***
## as.factor(grade)3  0.3751328  1.4551847  0.0705124  5.320 1.04e-07 ***
## nodes              0.0869173  1.0908065  0.0044943 19.339  < 2e-16 ***
## pgr               -0.0004244  0.9995757  0.0001200 -3.536 0.000406 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## chemo                1.1062     0.9040    0.9439    1.2964
## meno                 0.9787     1.0218    0.8056    1.1889
## hormon               0.9685     1.0325    0.8145    1.1516
## age                  1.0191     0.9812    1.0116    1.0268
## as.factor(grade)3    1.4552     0.6872    1.2674    1.6709
## nodes                1.0908     0.9168    1.0812    1.1005
## pgr                  0.9996     1.0004    0.9993    0.9998
##
## Concordance= 0.681  (se = 0.008 )
## Likelihood ratio test= 437.5  on 7 df,   p=<2e-16
## Wald test            = 552.3  on 7 df,   p=<2e-16
## Score (logrank) test = 606.8  on 7 df,   p=<2e-16
```

Table 8: Estimators and confidence intervals for relative risks under xxxxx model.

| | $exp(\hat{\beta})$ | 5% | 95% |
|---|---|---|---|
| Chemo | 1.052 | 0.896 | 1.235 |
| Meno | 1.073 | 0.881 | 1.307 |
| Hormon | 0.937 | 0.788 | 1.114 |
| Size (20-50) | 1.557 | 1.369 | 1.769 |
| Size (>50) | 2.276 | 1.902 | 2.722 |
| Age | 1.014 | 1.007 | 1.022 |
| Grade = 3 | 1.371 | 1.193 | 1.575 |
| Nodes | 1.076 | 1.065 | 1.086 |
| Pgr | 1.000 | 0.999 | 1.000 |
| Er | 1.000 | 1.000 | 1.000 |

1.0516048

Estimation and CI of relative risks for every pair of levels of the categorical variables are as below: HR of chemo vs non-chemo: 1.0423 with 95%CI [0.91353 1.18914] HR of meno vs non-meno: 0.8160 with 95%CI [0.69415 0.95916] HR of hormon treatment vs no treatment: 1.7815 with 95%CI [0.69415 0.95916] HR of size20-50 vs size<=20: 0.8650 with 95%CI [0.78503 0.95308] HR of size>50 vs size<=20: 0.9678 with 95%CI [0.79736 1.17457] HR of size>50 vs size20-50: exp(Beta(size>50)-Beta(size20-50))={exp(fit_ph4$coefficients['as.factor(size) > 50'])/$exp(fit_ph4$coefficients['as.factor(size)20-50'])}=1.118818 with 95%CI [??] HR of age i vs age i-1: 1.0075 with 95%CI [1.00110 1.01387] HR of grade 3 vs grade 2: 1.0575 with 95%CI [0.95655 1.16910] HR of recur vs non-recur: 0.0131 with 95%CI [0.01043 0.01646] HR of rtime i vs rtime i-1: 0.9985 with 95%CI [0.99841 0.99854]

```
##                   chisq df       p
## chemo             2.494  1 0.11425
## meno              4.257  1 0.03908
## hormon            0.548  1 0.45907
```

```
## age              14.055  1 0.00018
## as.factor(grade)  3.696  1 0.05455
## nodes             4.163  1 0.04132
## pgr              41.060  1 1.5e-10
## GLOBAL           58.534  7 3.0e-10
```

We run a statistical test based on Schoenfeld residuals for proportional hazard assumption for each covariate included in the cox fit. From the output, one can see that test is statistically significant for some covariates like size,grade, rtime and recur and so is the global test. So we can state that proportional hazard assumption is violated since there is significant dependency between Schoenfeld residuals and time.

##next step Q: should we also check outliers and nonlinearity between log hazard and covariates??

## 3.2 Parametric Regression Models

```
##   log(normal)       weibull   exponential log(logistic)
##      24274.47      24349.21      24456.48      24295.39

##          (Intercept)    as.factor(chemo)1 as.factor(size)20-50
##         9.0667707896        -0.0059565650        -0.4692326426
##   as.factor(size)>50    as.factor(grade)3   as.factor(hormon)1
##        -0.9483412908        -0.2626398851        -0.1847175598
##                  pgr                   er
##         0.0003415996        -0.0001415662

##                               2.5 %        97.5 %
## (Intercept)           8.9433004850  9.190241e+00
## as.factor(chemo)1    -0.1125498218  1.006367e-01
## as.factor(size)20-50 -0.5665404380 -3.719248e-01
## as.factor(size)>50   -1.0820595997 -8.146230e-01
## as.factor(grade)3    -0.3687207410 -1.565590e-01
## as.factor(hormon)1   -0.3136015894 -5.583353e-02
## pgr                   0.0001567376  5.264616e-04
## er                   -0.0002885341  5.401614e-06

##          (Intercept)    as.factor(chemo)1 as.factor(size)20-50
##        -1.188172e+01         7.805889e-03         6.149145e-01
##   as.factor(size)>50    as.factor(grade)3   as.factor(hormon)1
##         1.242771e+00         3.441812e-01         2.420665e-01
##                  pgr                   er
##        -4.476554e-04         1.855181e-04

##    as.factor(chemo)1 as.factor(size)20-50   as.factor(size)>50
##         0.0069872770         0.6100670434         1.2232404073
##    as.factor(grade)3   as.factor(hormon)1                  pgr
##         0.3369723661         0.2022551423        -0.0004441738
##                   er
##         0.0001822463
```

# 4 Next Steps for Aleks

- Keep researching to decide which test is appropriate for survival curves diff testing
- Format text
- Make all ugly R outputs into nice, coherent tables

# 5 References

Marta Sestelo. 2017. *A Short Course on Survival Analysis*. https://bookdown.org/sestelo/sa_financial/comparing-survival-curves.html.