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 ...

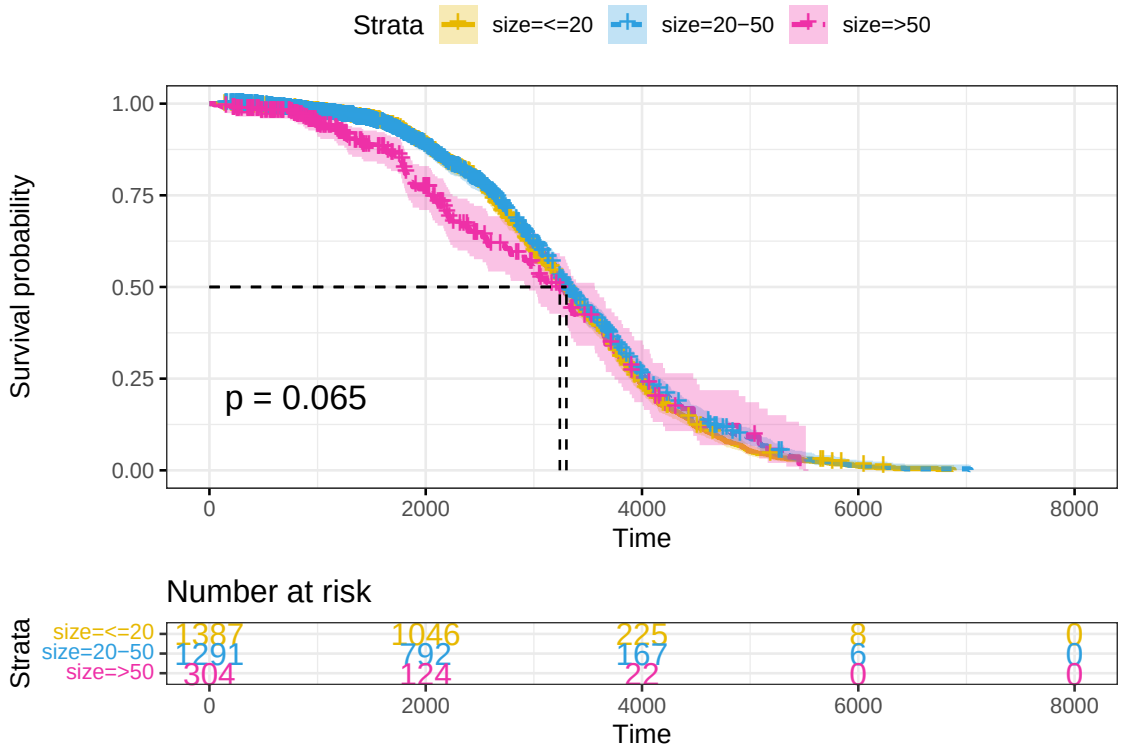
(a) For each of the levels of the categorical variable, compute the survival distribution. Plot them on the same graph. What do the graphs suggest?

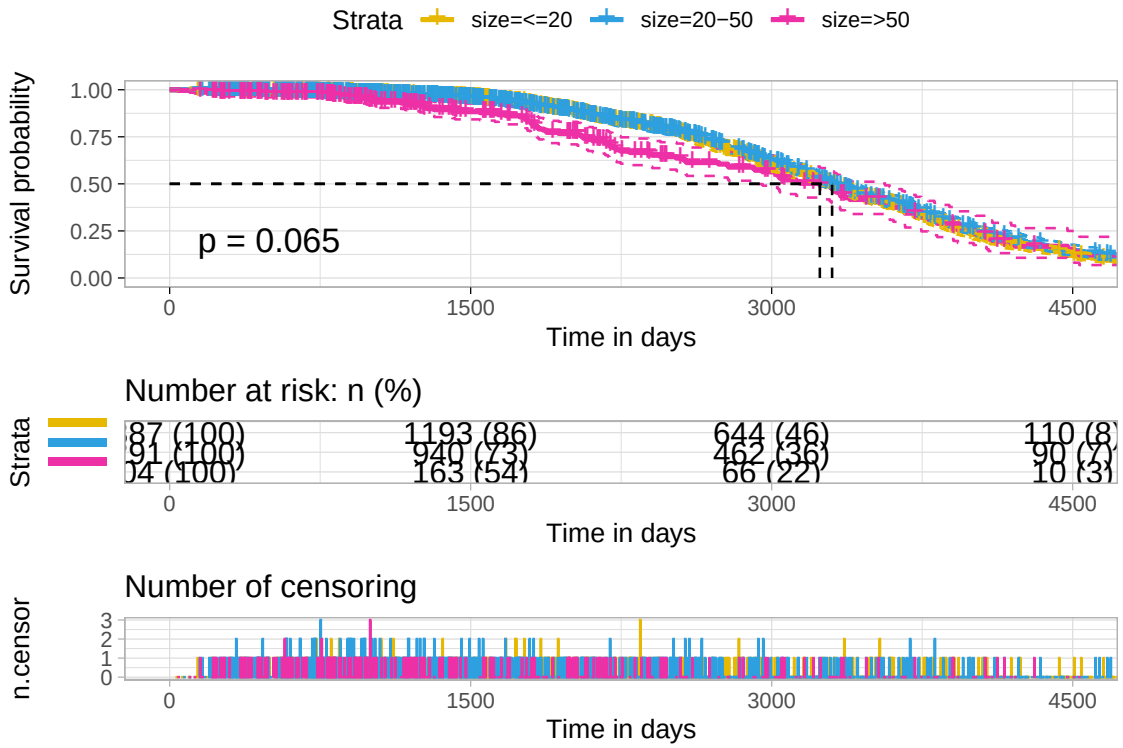
2.1.1 Size

```
## Call: survfit(formula = Surv(dtime, censored) ~ size, data = data)
##
##                n events median 0.95LCL 0.95UCL
## size=<=20   1387   1048   3289    3232    3370
## size=20-50  1291    734   3301    3230    3408
## size=>50     304    123   3240    2918    3565
```

```
##             records n.max n.start events    rmean se(rmean) median 0.95LCL
## size=<=20      1387  1387    1387   1048 3282.860  32.47471   3289    3232
## size=20-50     1291  1291    1291    734 3342.568  39.62310   3301    3230
## size=>50        304   304     304    123 3083.707 104.26023   3240    2918
##             0.95UCL
## size=<=20      3370
## size=20-50     3408
## size=>50       3565
```

```
##   time n.risk n.event n.censor      surv     upper     lower
## 1   36   1387       1        0 0.9992790 1.0000000 0.9978674
## 2   64   1386       1        0 0.9985580 1.0000000 0.9965631
## 3   97   1385       1        0 0.9978371 1.0000000 0.9953951
## 4  101   1384       1        0 0.9971161 0.9999422 0.9942980
## 5  129   1383       1        0 0.9963951 0.9995542 0.9932460
## 6  141   1382       0        1 0.9963951 0.9995542 0.9932460
```





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 2250, the probability of survival is approximately 0.625 for size>=50, and 0.85 for size<50. The median survival is approximately 3300 for size>=50, and a bit more for other two groups, suggesting 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.

The median survival times for each group can be seen from:

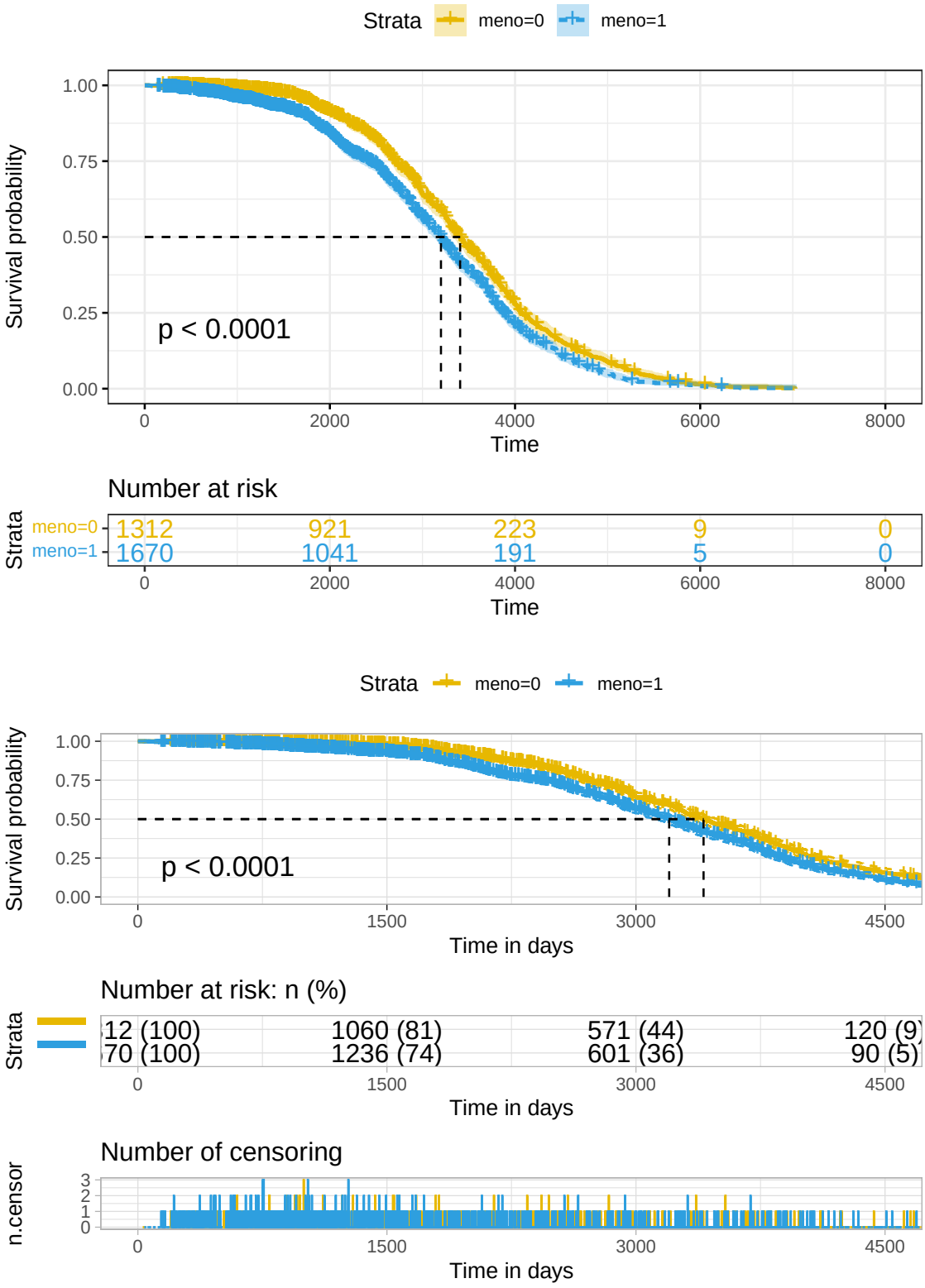
```
##             records n.max n.start events    rmean se(rmean) median 0.95LCL
## size=<=20      1387  1387    1387   1048 3282.860  32.47471   3289    3232
## size=20-50     1291  1291    1291    734 3342.568  39.62310   3301    3230
## size=>50        304   304     304    123 3083.707 104.26023   3240    2918
##             0.95UCL
## size=<=20      3370
## size=20-50     3408
## size=>50       3565
```

2.1.2 Menopause

```
## Call: survfit(formula = Surv(dtime, censored) ~ size, data = data)
##
##                n events median 0.95LCL 0.95UCL
## size=<=20   1387   1048   3289    3232    3370
```

```
## size=20-50 1291    734   3301    3230    3408
## size=>50    304    123   3240    2918    3565

##         records n.max n.start events    rmean se(rmean) median 0.95LCL 0.95UCL
## meno=0     1312  1312    1312    855 3451.798  35.59308   3407    3318    3487
## meno=1     1670  1670    1670   1050 3151.695  33.19115   3200    3112    3262
```



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 2250, the probability of survival is approximately 0.625 for size>=50, and 0.85 for size<50. The median survival is approximately 3300 for size>=50, and a bit more for other two groups, suggesting 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.

The median survival times for each group can be seen from:

```
##         records n.max n.start events    rmean se(rmean) median 0.95LCL 0.95UCL
## meno=0     1312  1312    1312    855 3451.798  35.59308   3407    3318    3487
## meno=1     1670  1670    1670   1050 3151.695  33.19115   3200    3112    3262
```

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.