Survival Analysis of German Breast Cancer

Group 3
Rui Li, Ruqian Qin
Jian Sun, Hanbing Zhou
The George Washington University
May 8, 2017

Abstract

The data set is with 686 patients from major clinical trials of the German Breast Cancer Study Group (GBSG) conducted between 1983 and 1989, of whom 299 had an event for recurrence-free survival and 171 died. The goals of this project are to build a prognostic model that predicts the clinical course of breast cancer patients, model the effects of standard prognostic factors, and evaluate efficacy of hormone therapy and provide treatment advice based on research. The Cox model is used to investigate the survival rate of the breast cancer patients and determine a final combination of the model with proper transformation of prognostic factors. The results reveal that the tumor size and tumor grade, the number lymph nodes and progesterone receptor number play significant role in patient survival. Other finding in this project includes: breast cancer patients aged over 55 at post menopause have higher survival rate; higher receptors elevate survival rate, while tumor grade increase; hormone therapy reduces the survival rate of patients in this dataset.

Keyword: ***Breast Cancer, Prognostic Factors, Cox Model, Hormone Therapy, Stratification***

1. Background, data set and research goals.

According to World Health Organization, "Breast cancer is the top cancer in women both in the developed and developing world". Breast cancer survival rates vary greatly worldwide, ranging from 80% or over in North America, Sweden and Japan to around 60% in middle-income countries and below 40% in low-income countries[1]. Since Breast cancer is highly curable if diagnosed at an early stage, for better prevention and treatment of breast cancer, a great deal of researches have been taken to detect its risk factors, like familial history of breast cancer, early menarche, late menopause, late age at first childbirth, using oral contraceptive or hormone replacement therapy, dietary effects, and shorter breastfeeding[2]. Even tumor size, tumor grade, histologic type, estrogen (ER) and progesterone receptor (PR) status, menopausal status and age, which are considered the prognostic value of standard factors, are still controversial among different studies[3]. Learning whether a tumor has estrogen and/or progesterone receptors helps doctors determine whether the cancer can be treated with hormone therapy.

The data is from major clinical trials of the German Breast Cancer Study Group (GBSG) conducted between 1983 and 1989. This Comprehensive Cohort Study (CCS) recruit all patients fulfilling the clinical eligibility criteria regardless of their consent to randomization [5]. The trials recruited 2084 patients by 124 centers in three clinical trials with different randomization rates [5]. The data we use is from one prospective study of node-positive breast cancer which they built a prognostic model that predicts the clinical course of breast cancer patients. The study investigated in 686 patients, of whom 299 had an event for recurrence-free survival and 171 died [3]. The standard prognostic factors are shown in Table 1(Appendix**),** and Figure 1 is the heatmap of Pearson Correlation for prognostic factors, which shows only age and menopause has relatively high correlation.

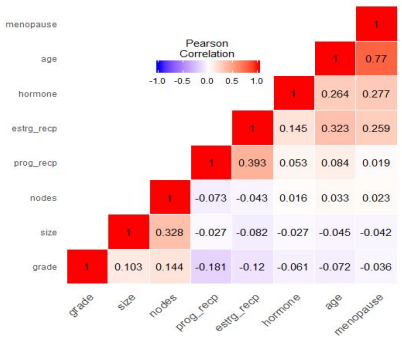


Figure 1. Correlation Heat Map

Our research goals are to build a prognostic model that predicts the clinical course of breast cancer patients, model the effects of standard prognostic factors and evaluate efficacy of hormone therapy and provide treatment advice based on research.

2. Data analysis, model selection and building, model diagnosis and problem solving

2.1 Survival curves and hypothesis test

Figure 2 is the K-M estimator for survival rate of recurrence and overall survival rate. Both curves are decreasing smoothly from beginning to the end of study.

Then we conducted hypothesis test on two groups based on different prognostic factors. Here Figure 3 shows the results for hormone therapy factor. P value of overall survival time(left) is larger than 0.05, indicating no significant difference, while small p value of recurrence free time(right) suggests significant difference between hormone therapy treated group and non treated group. And for menopause status, pre- and postmenopause groups show no significant difference for both survival outcomes, while for tumor grades and size, groups suggest significant difference.

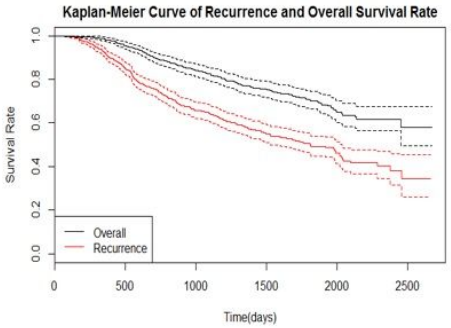


Figure 2. K-M estimator for survival rate

	Q	Var	Z	pNorm
1	-1.0166e+01	4.0347e+01	-1.60045	0.109498
n	-5.6160e+03	9.3689e+06	-1.83478	0.066539
sqrtN	-2.4085e+02	1.8250e+04	-1.78289	0.074605
S1	-9.2488e+00	2.9466e+01	-1.70383	0.088413
S2	-9.2448e+00	2.9312e+01	-1.70755	0.087721
FH_p=1_q=1	-8.1544e-01	7.2692e-01	-0.95642	0.338862

```
survdiff(formula = Surv(rectime, censrec) ~ hormone, data = gbcs,
    rho = 0)

             N Observed Expected (O-E)^2/E (O-E)^2/V
hormone=0 440      205      180      3.37      8.56
hormone=1 246       94      119      5.12      8.56

 Chisq= 8.6  on 1 degrees of freedom, p= 0.00343
```

Figure 3. Hypothesis test results for hormone therapy treated group and untreated group

2.2 Cox Model Selection

In this project, the Cox proportional hazard model was applied. We used stepwise model selection based on the AIC value.

```
coxph(formula = Surv(survtime, censdead) ~ size + grade + nodes +
    prog_recp, data = gbcs)

  n= 686, number of events= 171

               coef exp(coef)  se(coef)      z Pr(>|z|)
size       0.013481  1.013572  0.004726  2.853  0.00433 **
grade2     0.791057  2.205727  0.425267  1.860  0.06287 .
grade3     1.150121  3.158574  0.441056  2.608  0.00912 **
nodes      0.051290  1.052628  0.009495  5.402 6.59e-08 ***
prog_recp -0.005349  0.994665  0.001152 -4.644 3.41e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
size          1.0136     0.9866    1.0042    1.0230
grade2        2.2057     0.4534    0.9584    5.0762
grade3        3.1586     0.3166    1.3307    7.4975
nodes         1.0526     0.9500    1.0332    1.0724
prog_recp     0.9947     1.0054    0.9924    0.9969
```

```
coxph(formula = Surv(rectime, censrec) ~ hormone + size + grade +
    nodes + prog_recp, data = gbcs)

  n= 686, number of events= 299

               coef  exp(coef)   se(coef)      z Pr(>|z|)
hormone    0.3235201  1.3819839  0.1258239  2.571   0.0101 *
size       0.0073129  1.0073397  0.0038897  1.880   0.0601 .
grade2     0.6440346  1.9041479  0.2490184  2.586   0.0097 **
grade3     0.7882290  2.1994977  0.2682585  2.938   0.0033 **
nodes      0.0489976  1.0502178  0.0074529  6.574 4.89e-11 ***
prog_recp -0.0022168  0.9977856  0.0005538 -4.003 6.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
hormone       1.3820     0.7236    1.0799    1.7685
size          1.0073     0.9927    0.9997    1.0150
grade2        1.9041     0.5252    1.1688    3.1022
grade3        2.1995     0.4546    1.3001    3.7211
nodes         1.0502     0.9522    1.0350    1.0657
prog_recp     0.9978     1.0022    0.9967    0.9989
```

Figure 4. Models selection of overall survival (left) and recurrence-free survival model (right)

For overall survival, the size and grade 3 of tumor, the nodes, PR concentrations are significant factors.

Also, we could obtain some information:

1. given the same conditions, as the tumor size or positive lymph nodes increase, the chance of having cancer recurrence increased.
2. given the same body conditions, the relative risk (RR) for tumor grade 2 v.s tumor grade 1 is around 2.21. Similarly, for RR for tumor grade 3 v.s grade 1 is 3.16. The higher tumor grade, the higher chance of death;
3. given the same conditions, the higher PR concentration, the lower the probability of cancer death. This also will be discussed more later in this report.

For recurrence model, beside all the factors in overall survival model, it also includes the hormone therapy. Interestingly, given the same conditions, patients with hormone therapy show 0.38 higher in chance of having cancer recurrence, and we will discuss the efficacy of this treatment later.

2.3 Model Diagnosis

After selecting the Cox model, the further diagnosis is necessary. The diagnosis includes testing proportional assumption, analyzing the Cox-Snell residual plot, transforming the predictors, and other diagnosis.

a. overall survival model

First, for the purpose of testing the proportionality, the result of using the interaction of all the predictor with the transformation of time were produced:

A p-value less than 0.05 indicates a possible violation of proportional assumption. In the overall survival, the PR concentration (prog_recp) might be considered as a time-dependent-covariate. In this case, two common approaches: creating a time-dependent-covariate in long format; stratifying that covariate. We applied the first approach by using "Survsplit" function, however, the Cox-Snell residual plot showed poor performance of the model. So that we choose to the other approach that stratified the PR concentration into three levels as we mentioned in before, and ran the log-likelihood-ratio test that there is no difference between the stratified model and the separated.

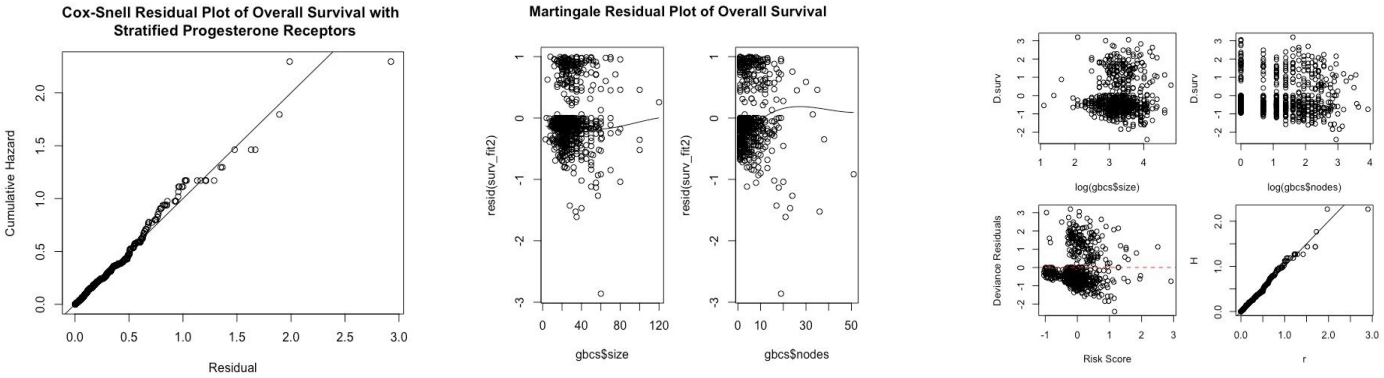


Figure 5. Cox-Snell Residual Plot (left), Martingale Residual Plot (middle) and Residual Plot of Transformed Model (right)

The Cox-Snell residual plot shows that most the data points are roughly along the line, but still several dots are away from it. Then, from the Martingale residuals, the residual of variables nodes and size are not random pattern, the transformation might be taken into thought.

After the log-transformation, we applied the deviance plot to test the goodness-of-fit (**Figure 5. right)**. The top two plots are the deviance v.s transformed nodes and size, which are randomly distributed. From the deviance plot of risk score, the model is fairly good-of-fit, but there are still large number of negative residual, which might because of the large number of censoring data or Cox model might not be the good method for this case. Also, from the Cox-Snell residuals plot, the model performed slightly better than non-transformed model.

b. recurrence-free survival model

Similar to overall survival model diagnosis, from the proportional assumption test, tumor grade3 is the one might be considered as a time-dependent-covariate, and the stratified model with tumor grade was applied in this matter.

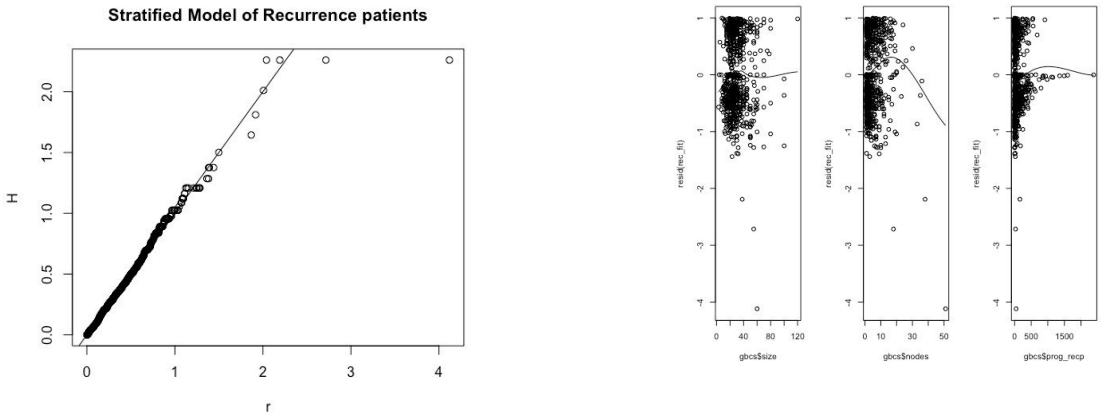


Figure 6. Cox-Snell Residual Plot (left) and Martingale Residual Plot (right) After Stratification

The Cox-Snell residual of stratified plot on the left and Martingale residual plot on the right suggest the variables transformation is needed.

Like the overall survival model, the log-transformation of nodes was used. Also from the paper of Sauerbrei etc., they compared different transformation of standard prognostic factor of breast cancer, and recommended to use square root transformation (the power of 0.5) has the best performance[7]. In that we applied the similar approach to our case.

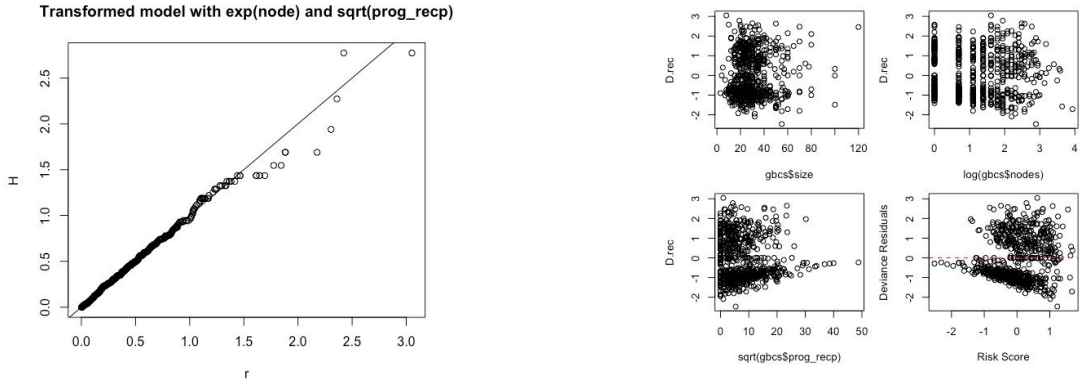


Figure 7. Cox-Snell Residual Plot (left) and Martingale Residual Plot (right) After Stratification

The Cox-Snell plot on the left shows the model is good-of-fit. and the deviance plot against each of the predictor has random pattern. Meanwhile the deviance plot of risk score indicates similar pattern like overall survival, and it still could consider the model is fairly good in the performance.

2.4 Specific Problems Solving

The stratified proportional hazard model is used to solve the following three problems.

2.4.1 What is the effect of late menopause on patients who are older than 55?

According to the medical knowledge, the menopause status has big effect on breast cancer. To verify this statement, several comparisons are done. The controlled variables are age (over 55) and tumor grade (1, 2, 3). The following are the steady survival rate for different status.

	Overall Survival	Recurrence-free Surviva

	Postmenopause	Premenopause	Postmenopause	Premenopause
Grade 1	0.911	0.848	0.734	0.602
Grade 2	0.728	0.57	0.486	0.303
Grade 3	0.555	0.35	0.383	0.203

For overall survival part, in each grade, increasing time reduces the survival rate and enlarges the difference between post and pre menopause. In a meanwhile, increasing tumor grade obviously lower the survival rate, but in each grade, the survival rate of postmenopause is higher than that of pre menopause.

For recurrence part, the similar results are observed. Figure 1 and 2 are the detailed plots for this problem.

Thus, the conclusions for this problem is that higher tumor grade leads to less survival rate; the breast cancer patients, who aged over 55, have more probability to survival if they are at post menopause.

2.4.2 Does Progesterone Receptor (PR) or Estrogen Receptor (ER) work well on curbing Breast Cancer?

According to the medical knowledge, the Receptors on the cancer cell would combine with progesterone and estrogen and restrain the growing of cancer cell. To verified this statement, several comparisons are done. The controlled variables are age (over 55) and tumor grade (2, 3).

The following tables are the steady survival rate for different status.

Receptor Concentration	Overall Survival				Recurrence-free Survival			
	PR		ER		PR		ER	
	Grade 2	Grade 3	Grade 2	Grade 3	Grade 2	Grade 3	Grade 2	Grade 3
More than 90	0.806	0.758	0.697	0.562	0.514	0.499	0.347	0.276
21 to 90	0.633	0.556	0.642	0.492	0.378	0.365	0.393	0.324
Less than 20	0.366	0.276	0.433	0.264	0.137	0.128	0.266	0.203

For overall survival part, in each receptor number, increasing tumor grade reduces the survival rate. In a meanwhile, receptor number has positive relationship with survival rate. Both PR and ER have the same results.

For recurrence part, the similar results are observed in PR. About ER, higher tumor grade leads to smaller survival rate as well. And as receptor number increases, the survival rate improves in most of time, but in the end, the order of survival rate is 21 to 90, more than 90 and less than 20 from high to low. Figure 3 and 4 are the detailed plots for this problem.

Thus, the conclusions for this problem is that higher receptors repress breast cancer and elevate survival rate; both PR and ER study lead to same result as tumor grade increasing.

2.4.3 Is it useful for patients to control breast cancer by hormone therapy?

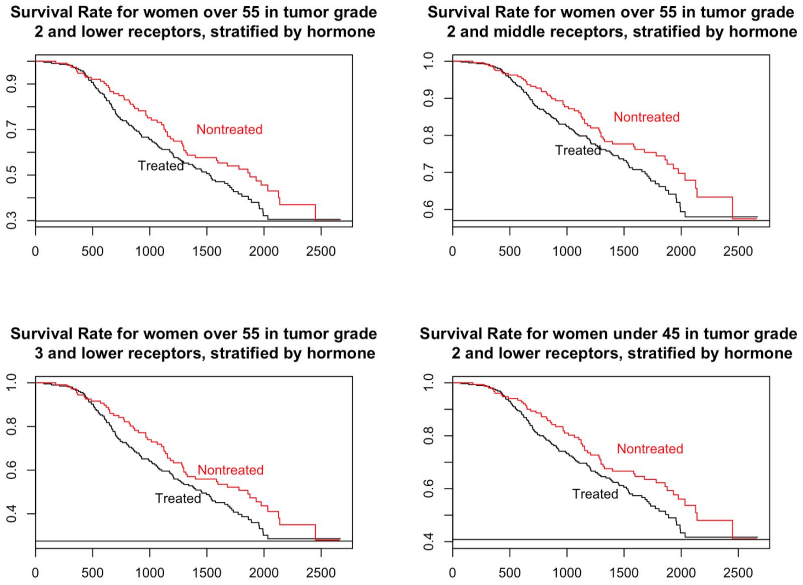
Currently, the hormone therapy is a popular choice on curing breast cancer, but it also has some side effects or even negative effects on patients. To verify that if hormone therapy works well for patients in this dataset, several comparisons are done. The controlled variables are age (under 45, over 55), PR (0 to 20, 21 to 90), ER (0 to 20, 21 to 90) and tumor grade (2, 3).

For overall survival results, after setting age is over 55, tumor grade is 2, we changed receptor number from lower (0 to 20) to middle (21 to 90), we found that the survival rate is reducing as the time increases, and in most of time, the middle receptor number does not make hormone therapy work better than the lower one, and in the end, the steady survival rate is nearly same for two status. So higher PR number increases the survival rate, but does not make hormone therapy work better. However, tumor grade or age may affect the performance of hormone therapy as well. Therefore, we do comparison one by one.

After setting age is over 55, PR number is low, we changed tumor grade from 2 to 3, we found that the survival rate is reducing as the time increases, and in most of time, the serious tumor grade does not make hormone therapy work better than the lower one, and in the end, the steady survival rate is nearly same for two status.

Next,after setting tumor grade is 3, PR number is low, we changed age from over 55 to 45, because we guess that maybe younger women have more power to heal under hormone therapy. However, we found that the survival rate is reducing as the time increases, and in most of time, the younger age does not make hormone therapy work better than the lower one, and in the end, the steady survival rate is nearly same for two status. Thus serious tumor grade and younger age don't make hormone therapy work better.

Figure 8. the plots for different status.



What's more, the results don't change after changing tumor grade to tumor size. And for recurrence part, the similar results are observed. Figure 5 is another detailed plots for this problem.

Hence, the conclusions for this problem is drawn that on different conditions, getting hormone therapy reduces the survival rate of breast cancer patients in this dataset.

3. Conclusions and discussion

There are hundreds of papers and projects in the last 2 decades, only the nodal status is commonly accepted by as a strong factor in breast cancer patients without metastases (M0)[3]. While other prognostic factors such as tumor size, tumor grade, histologic type, estrogen (ER) and progesterone receptor (PR) status, menopausal status and age is still controversial, recently many new controversial factors (clinicopathological, biological, molecular) have been proposed and investigated[3]. In one study, the pre-menopausal node-positive patients have higher risk of early recurrence compared to postmenopausal patients [6]. But in our result, the menopause status is not a significant prognostic factor when we do

the hypothesis test and modelling. And the tumor grade and tumor size are significant prognostic factors for breast cancer. For final model of overall survival time, the tumor size and grade 3 of tumor, the nodes, PR concentrations are the significant prognostic factors we choose.

The Cox model is commonly used in oncology study[7]. In this project, the Cox model sufficiently predicts the survival of breast cancer patients. But the asymmetrical deviance plot might imply still some room for improvement. For example, the proportional assumption is not applicable, other parametric method, such as exponential model might be a good candidate.

Aiming to question 1, the conclusion is that the breast cancer patients, who aged over 55, have more probability to survival if they are at post menopause. Thus we suggest that patients should inquire if pre menopause over 55, otherwise hard to elevate the survival rate for themselves.

Aiming to question 2, the conclusion is that higher receptors elevate survival rate; both PR and ER study lead to same result as tumor grade increasing. Thus we suggest that for receptor-test result positive patients, the hormone therapy is recommended, otherwise they can do chemotherapy.

Aiming to question 3, the conclusion is that getting hormone therapy reduces the survival rate of patients in this dataset. Thus we suggest that patients should check if the hormone therapy has efficacy on themselves before taking it.

There still are many limitations of our study. From statistical aspects, inadequate sample sizes(we only have 686 patients), inadequate use of statistical methods(our Cox model is still not perfect fit for the data set) and difficulties in comparing multivariable models with different factors or different categorizations of the factors(there are still more models we can try to fit and compare). And also the heterogeneity in patient populations and treatment or limitations of follow-up often cause difficulty for model fit and analysis.

4. Roles

Rui Li: Background and variables exploration, heatmap, final report writing
Ruqian Qin: Proposal writing, Model diagnosis and final report writing
Jian Sun:Model building, problems solving, final report writing
Hanbing Zhou: Background and conclusion, hypothesis test, KM survival curve, final report writing

5. References:

1. Coleman MP1, Quaresma M, etc. Cancer survival in five continents: a worldwide population-based study (CONCORD).Lancet Oncol. 2008 Aug;9(8):730-56.
2. "Breast Cancer: Prevention and Control." World Health Organization. World Health Organization, n.d. Web. 23 Apr. 2017.vbk
3. W Sauerbrei, P Royston, etc. Modelling the effects of standard prognostic factors in node-positive breast cancer. Br J Cancer. 1999 Apr; 79(11-12): 1752–1760.
4. Schmoor C1, Olschewski M, Schumacher M. Randomized and nonrandomized patients in clinical trials: experiences with comprehensive cohort studies. Stat Med. 1996 Feb 15;15(3):263-71.
5. Romano Demicheli, Gianni Bonadonna etc. Menopausal status dependence of the timing of breast cancer recurrence after surgical removal of the primary tumor. Breast Cancer Research 20046:R689
6. Sauerbrei, W., Royston, P., Bojar, H., Schmoor, C., Schumacher, M., & German Breast Cancer Study Group. (1999). Modelling the effects of standard prognostic factors in node-positive breast cancer. British Journal of Cancer, 79(11-12), 1752.
7. Cox D.R (1972) Regression models and life tables (with discussion). J R Stat Soc B 34: 187–220.

6. Appendix (additional results and R Codes):

Tables and Figures

Table 1. Main variables investigated in German cancer study.

Variable	Codes/Values	Mean	Sd	Quartiles		
Numeric:				25%	50%	75%
Age at diagnosis	Years	53.05	10.12	46	53	61
Tumor Size	mm	29.33	14.30	20	25	35
# of Nodes involved	1-51	5.01	5.48	1	3	7
# of Progesterone Receptors	0-2380	110.00	202.33	7	33	132
# of Estrogen Receptors	0-1144	96.25	152.08	8	36	114
Categorical:			Codes: Number			
Menopausal status	1=Yes, 2=No		Yes: 290; No:396			
Hormone Therapy	1=Yes; 2=No		Yes: 440; No: 246			
Tumor Grade	1-3(G-1, G-2, G-3)		G-1:81; G-2:444; G-3:161			

Figure 1. Survival Rate for Menopause Status in survival part

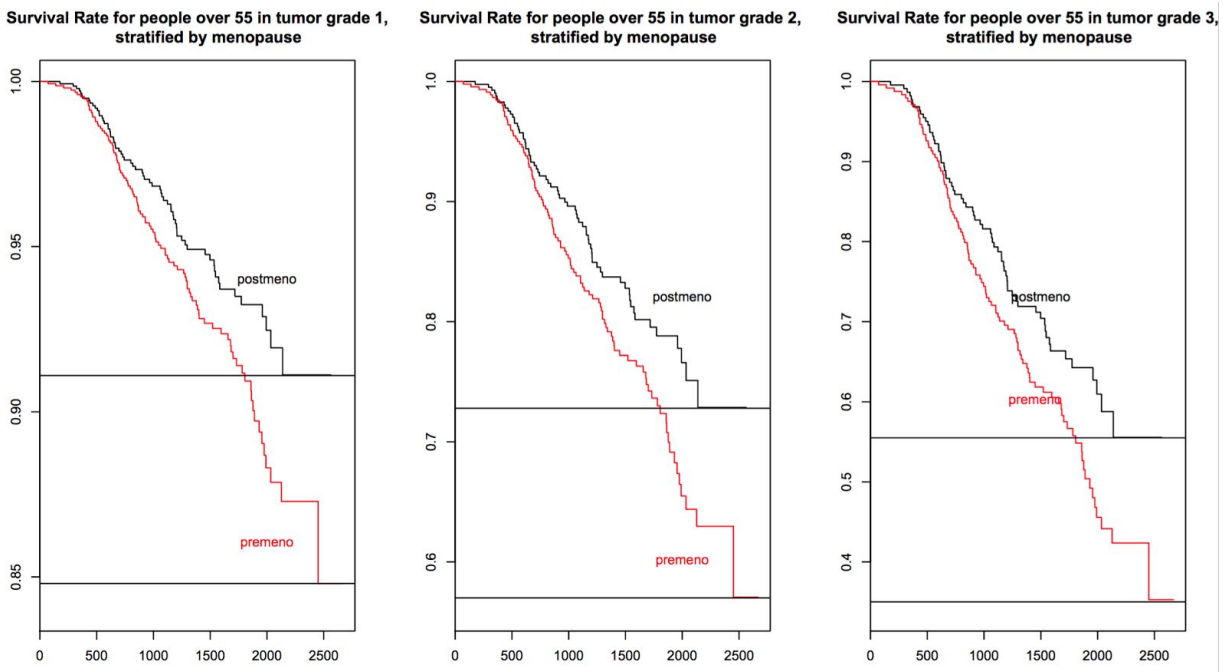


Figure 2. Survival Rate for Menopause Status in recurrence part

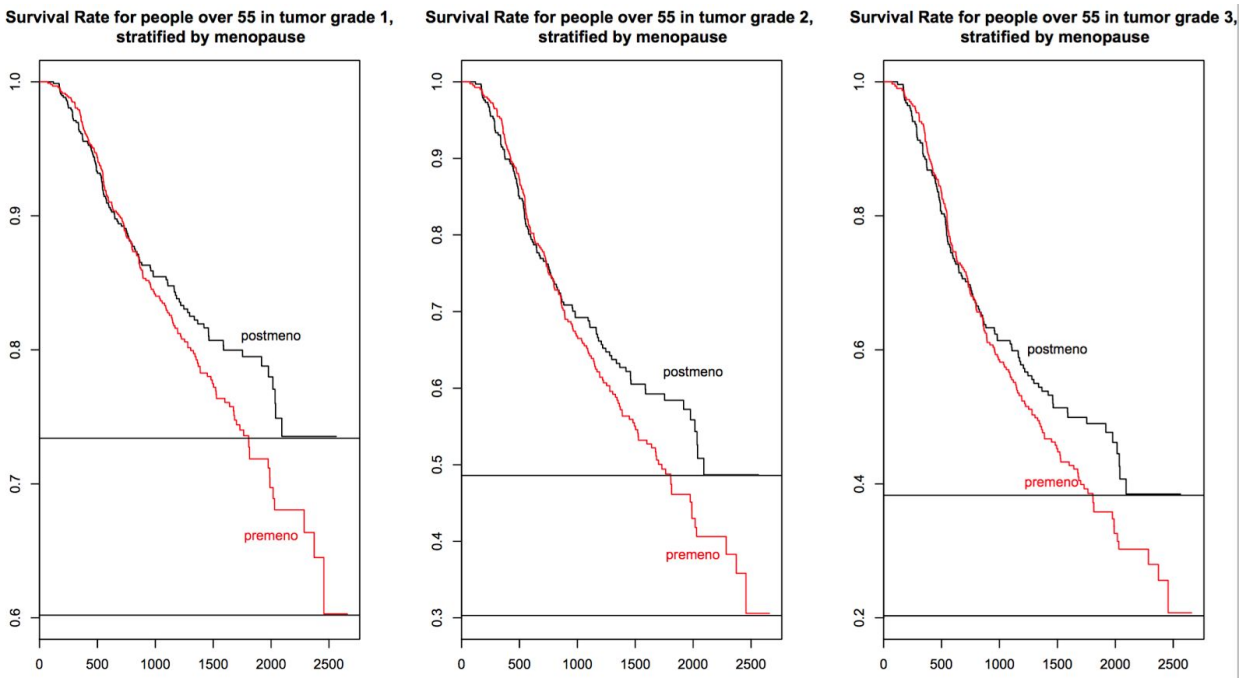


Figure 3. Survival Rate for Two Kinds of Receptors in Survival Part

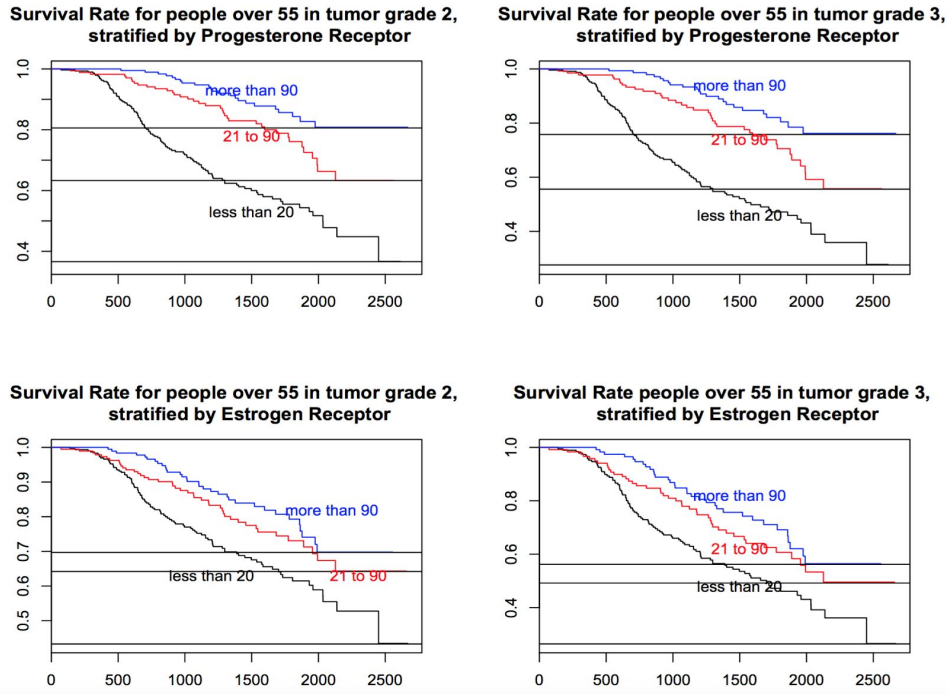


Figure 4. Survival Rate for Two Kinds of Receptors in Recurrence Part

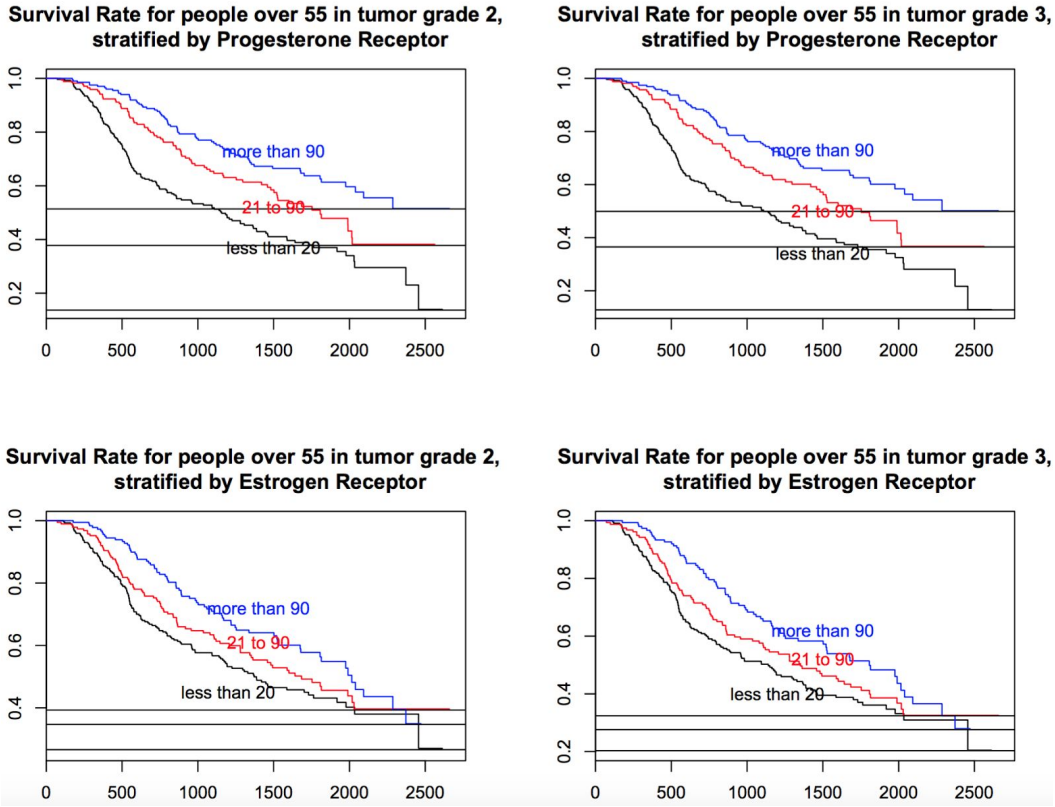
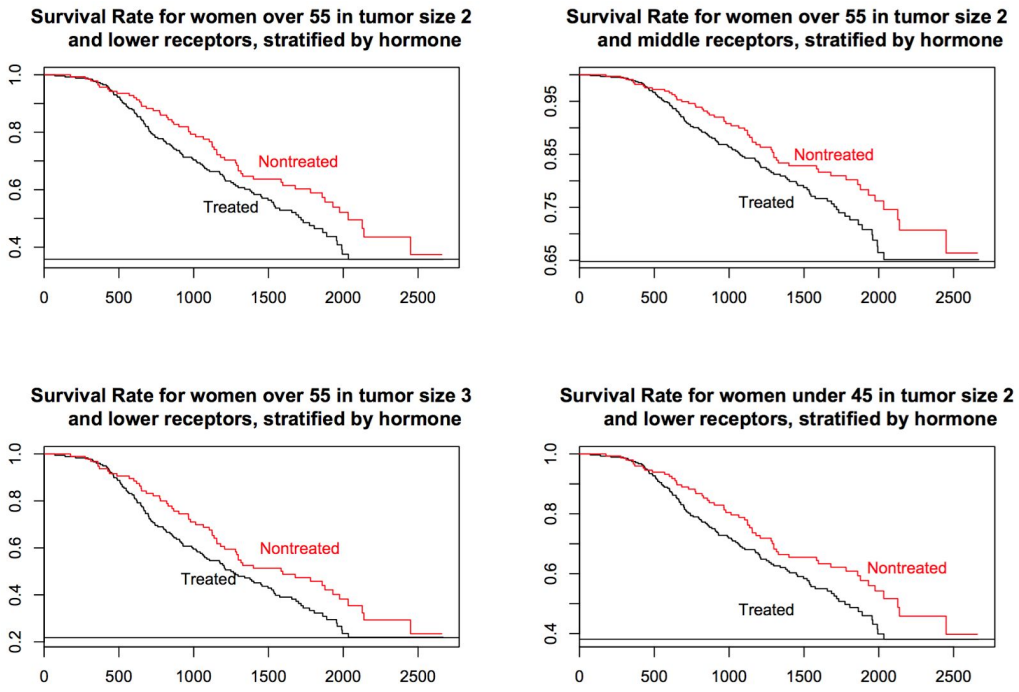


Figure 5. Survival Rate for Tumor Size in Survival Part



R Code

```
##load library and data
library(XLConnect)
library(KMsurv)
library(survival)
library(OIsurv)
library(survMisc)
library(MASS)
library(dplyr)
library(car)
wb = loadWorkbook("gbcs.xls")
gbcs = readWorksheet(wb, sheet = "GBCS",header = TRUE)

##Data cleaning
gbcs[,"menopause"] <- ifelse(gbcs[,"menopause"] == 2, 0, 1)
gbcs[,"hormone"] <- ifelse(gbcs[,"hormone"] == 2, 0, 1)
gbcs[,"grade"]=as.factor(gbcs$grade)

##KM and NA estimator
kMfit_rec = survfit(Surv(rectime, censrec)~1, data=gbcs,
type="kaplan-meier")
kMfit_dead = survfit(Surv(survtime, censdead)~1, data=gbcs,
type="kaplan-meier")
NAfit_dead = survfit(Surv(survtime, censdead)~1, data=gbcs,
type="fl")
NAfit_rec = survfit(Surv(rectime, censrec)~1, data=gbcs,
type="fl")

#plot KM for recurrunce and overall
plot(kMfit_rec, xlab="Time(days)", ylab="Survival Rate",
col="red")
lines(kMfit_dead, col="black")
legend("bottomleft", c("Overall","Recorrence"),
col=c("black", "red"), lty=c(1,1))
title("Kaplan-Meier Curve of Recurrence and Overall Survival
Rate")

##Hypothesis Testing
#1) menopause
plot(survfit(Surv(survtime, censdead)~menopause, data=
gbcs))
meno_os <- ten(survfit(Surv(survtime,
censdead)~menopause, data= gbcs))
comp(meno_os)
plot(survfit(Surv(rectime, censrec)~menopause, data= gbcs))
meno_rec <- ten(survfit(Surv(rectime, censrec)~menopause,
data= gbcs))
comp(meno_rec)

#2) hormone
plot(survfit(Surv(survtime, censdead)~hormone, data= gbcs))
horm_os <- ten(Surv(survtime, censdead)~hormone, data=
gbcs)
comp(horm_os)
plot(survfit(Surv(rectime, censrec)~hormone, data= gbcs))
horm_rec <- ten(Surv(rectime, censrec)~hormone, data=
gbcs, rho = 0)
comp(horm_rec)
plot(survfit(Surv(survtime, censdead)~grade, data= gbcs))
grade_os <- survdiff(Surv(survtime, censdead)~grade, data=
gbcs,rho = 0)
print(grade_os)
plot(survfit(Surv(rectime, censrec)~grade, data= gbcs))
grade_rec <- ten(Surv(rectime, censrec)~grade, data=
gbcs,rho = 0)
comp(grade_rec)
```

```
## Model Selection
rec_fit<-coxph(Surv(rectime,censrec)~age+menopause+horm
one+size+grade+nodes+prog_recp+estrg_recp, data=gbcs)
summary(rec_fit)
surv_fit<-coxph(Surv(survtime,censdead)~age+menopause+h
ormone+size+grade+nodes+prog_recp+estrg_recp,
data=gbcs)
summary(surv_fit)
selection_surv=stepAIC(surv_fit)
selection_rec=stepAIC(rec_fit)
summary(selection_surv)
summary(selection_rec)

##Model Diagnosis
#Test the proportional assumption on recurrence patients:
propcheck_rec=cox.zph(selection_rec, transform="km",
global=TRUE)
plot(propcheck_rec)
print(propcheck_rec)
rec_fit<-coxph(Surv(rectime,censrec)~hormone+size+strata(g
rade)+nodes+prog_recp,data=gbcs)
summary(rec_fit)
```

```
model.sep.1<-coxph(Surv(rectime,censrec)~hormone+size+
nodes+prog_recp, data=gbcs, subset=gbcs[,"grade"]==1)
model.sep.2<-coxph(Surv(rectime,censrec)~hormone+size+
nodes+prog_recp, data=gbcs,subset=gbcs[,"grade"]==2)
model.sep.3<-coxph(Surv(rectime,censrec)~hormone+size+
nodes+prog_recp, data=gbcs,subset=gbcs[,"grade"]==3)
LL.sep = model.sep.1$loglik[2] + model.sep.2$loglik[2] +
model.sep.3$loglik[2]
LL.strat = rec_fit$loglik[2]
chisq.t = 2*(LL.sep - LL.strat)
chisq.t > qchisq(0.95, 8)
rec_diff=gbcs$censrec-resid(rec_fit)
rec_haz=survfit(Surv(rec_diff,gbcs$censrec)~1,type="fl")
plot(rec_haz$time,-log(rec_haz$surv),ylab="H",xlab="r",
main="Stratified Model of Recurrence patients")
abline(0,1)
par(mfrow=c(1,3))
scatter.smooth(gbcs$size,resid(rec_fit))
scatter.smooth(gbcs$nodes,resid(rec_fit))
scatter.smooth(gbcs$prog_recp,resid(rec_fit))
rec_fit2<-coxph(Surv(rectime,censrec)~hormone+size+strata(
grade)+
I(log(nodes))+I(sqrt(prog_recp)),data=gbcs)
summary(rec_fit2)
rec_diff2=gbcs$censrec-resid(rec_fit2)
rec_haz2=survfit(Surv(rec_diff2,gbcs$censrec)~1,type="fl")
plot(rec_haz2$time,-log(rec_haz2$surv),ylab="H",xlab="r",
main="Transformed model with exp(node) and
sqrt(prog_recp)")
abline(0,1)
D.rec = resid(rec_fit2, type="deviance")
par(mfrow=c(2,2),mar=c(4,4,1,1), oma=c(1,1,3,1))
plot(D.rec~gbcs$size)
plot(D.rec~log(gbcs$nodes))
plot(D.rec~sqrt(gbcs$prog_recp))
plot(rec_fit2$linear.predictors, D.rec,xlab="Risk
Score",ylab="Deviance Residuals")
abline(0,0,lty=2,col='red')

propcheck_surv=cox.zph(selection_surv, transform="km",
global=TRUE)
plot(propcheck_surv)
print(propcheck_surv)
gbcs$pro.stra=recode(gbcs$prog_recp, "lo:20=1; 21:90=2 ;
91:hi=3")
surv_fit2 <-
coxph(Surv(survtime,censdead)~size+grade+nodes+
strata(pro.stra),data=gbcs)
propcheck_surv_tdc <- cox.zph(surv_fit2)
plot(propcheck_surv_tdc)
print(propcheck_surv_tdc)
summary(surv_fit2)
model.sep.1<-coxph(Surv(rectime,censrec)~size+factor(grade
)+nodes, data=gbcs, subset=gbcs[,"pro.stra"]==1)
model.sep.2<-coxph(Surv(rectime,censrec)~size+factor(grade
)+nodes, data=gbcs,subset=gbcs[,"pro.stra"]==2)
model.sep.3<-coxph(Surv(rectime,censrec)~size+factor(grade
)+nodes, data=gbcs,subset=gbcs[,"pro.stra"]==3)
LL.sep = model.sep.1$loglik[2] + model.sep.2$loglik[2] +
model.sep.3$loglik[2]
LL.strat = surv_fit2$loglik[2]
chisq.t = 2*(LL.sep - LL.strat)
chisq.t > qchisq(0.95, 8)
surv_diff=gbcs$censdead-resid(surv_fit2)
surv_haz=survfit(Surv(surv_diff,gbcs$censdead)~1,type="fl")
plot(surv_haz$time,-log(surv_haz$surv),ylab="Cumulative
Hazard",xlab="Residual")
title("Cox-Snell Residual Plot of Overall Survival with \n
Stratified Progesterone Receptors")
abline(0,1)

par(mfrow=c(1,2),mar=c(4,4,1,1), oma=c(1,1,3,1))
scatter.smooth(gbcs$size,resid(surv_fit2))
scatter.smooth(gbcs$nodes,resid(surv_fit2))
title("Martingale Residual Plot of Overall Survival", outer = T)
surv_fit3 <- coxph(Surv(survtime,censdead)~I(log(size))+
grade+I(log(nodes))+strata(pro.stra),data=gbcs)
surv_diff=gbcs$censdead-resid(surv_fit3)
surv_haz=survfit(Surv(surv_diff,gbcs$censdead)~1,type="fl")
plot(surv_haz$time,-log(surv_haz$surv),ylab="H",xlab="r")
abline(0,1)
D.surv = resid(surv_fit3, type="deviance")

par(mfrow=c(2,2), mar=c(4,4,1,1), oma=c(1,1,3,1))
plot(D.surv~log(gbcs$size))
plot(D.surv~log(gbcs$nodes))
plot(surv_fit2$linear.predictor, D.surv,xlab="Risk
Score",ylab="Deviance Residuals")
abline(0,0,lty=2,col='red')
plot(surv_haz$time,-log(surv_haz$surv),ylab="H",xlab="r")
abline(0,1)
propcheck_surv=cox.zph(surv_fit3, transform="km",
global=TRUE)
print(propcheck_surv)
```

##Specific Problems Solving

#total categorial

```
gbcs[,"grade"]=as.factor(gbcs$grade)
gbcs[,"menopause"]=as.factor(gbcs$menopause)
gbcs[,"hormone"] =as.factor(gbcs$hormone)
gbcs$age.stra=recode(gbcs$age, "lo:45=1; 46:55=2; 55:hi=3")
gbcs$size.stra=recode(gbcs$size, "lo:20=1; 21:30=2; 31:hi=3")
gbcs$nodes.stra=recode(gbcs$nodes, "lo:3=1; 4:9=2;
10:hi=3")
gbcs$est.stra=recode(gbcs$estrg_recp, "lo:20=1; 21:90=2 ;
91:hi=3")
```

#1 The efficient variables that determined the survival rate for breast cancer patients.

 # For survival part

```
par(mfrow=c(1,3))
surv_fit<-coxph(Surv(survtime,censdead)~age.stra+
strata(menopause)+as.factor(grade), data=gbcs)
summary(surv_fit)
#grade 1
tp1=survfit(surv_fit,
newdata=data.frame(grade=1,age.stra=3))
plot(tp1, ylim=c(0.84,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1,7),
   main="Survival Rate for people over 55 in tumor grade 1,
   stratified by menopause")
abline(0.848,0)
abline(0.911,0)
text(2000,.94,"postmeno", col="black")
text(2000,.86,"premeno", col="red")

#grade 2
tp2=survfit(surv_fit,
newdata=data.frame(grade=2,age.stra=3))
plot(tp2, ylim=c(0.56,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),main="Survival Rate for people over 55 in
tumor grade 2, stratified by menopause")
abline(0.57,0)
abline(0.728,0)
text(2000,.82,"postmeno", col="black")
text(2000,.6,"premeno", col="red")
#grade 3
tp3=survfit(surv_fit,
newdata=data.frame(grade=3,age.stra=3))
plot(tp3,ylim=c(0.34,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 3,
   stratified by menopause")
abline(0.35,0)
abline(0.555,0)
text(1500,.73,"postmeno", col="black")
text(1450,.6,"premeno", col="red")
```

#For Recurrence Part

```
par(mfrow=c(1,3))
rec_fit<-coxph(Surv(rectime,censrec)~age.stra+strata(menop
ause)+
            as.factor(grade), data=gbcs)
#grade 1
tp1=survfit(rec_fit,
newdata=data.frame(grade=1,age.stra=3))
plot(tp1, ylim=c(0.6,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1,7),
   main="Survival Rate for people over 55 in tumor grade 1,
   stratified by menopause")
abline(0.602,0)
abline(0.734,0)
text(2000,.81,"postmeno", col="black")
text(2000,.66,"premeno", col="red")

#grade 2
tp2=survfit(rec_fit,
newdata=data.frame(grade=2,age.stra=3))
plot(tp2, ylim=c(0.3,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 2,
   stratified by menopause")
abline(0.303,0)
abline(0.486,0)
text(2000,.62,"postmeno", col="black")
text(2000,.38,"premeno", col="red")
```

```
#grade 3
tp3=survfit(rec_fit,
newdata=data.frame(grade=3,age.stra=3))
plot(tp3,ylim=c(0.2,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 3,
   stratified by menopause")
abline(0.203,0)
abline(0.383,0)
text(1500,.6,"postmeno", col="black")
text(1450,.4,"premeno", col="red")
```

```
## 2 Does PR and ER work well on curbing cancer
#for overall survival
#for Progesterone
par(mfrow=c(2,2))
surv_fit<-coxph(Surv(survtime,censdead)~age.stra+as.factor(
grade)
          +strata(pro.stra), data=gbcs)
tpa=survfit(surv_fit,
newdata=data.frame(grade=2,age.stra=3))
plot(tpa, ylim=c(0.35,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 2,
   stratified by Progesterone Receptor")
abline(0.366,0)
abline(0.633,0)
abline(0.806,0)
text(1500,.53,"less than 20", col="black")
text(1500,.78,"21 to 90", col="red")
text(1500,.93,"more than 90", col="blue")

tpb=survfit(surv_fit,
newdata=data.frame(grade=3,age.stra=3))
plot(tpb, ylim=c(0.27,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 3,
   stratified by Progesterone Receptor")
abline(0.276,0)
abline(0.556,0)
abline(0.758,0)
text(1500,.46,"less than 20", col="black")
text(1500,.74,"21 to 90", col="red")
text(1500,.94,"more than 90", col="blue")

#for Estrogen
#par(mfrow=c(1,2))
surv_fit<-coxph(Surv(survtime,censdead)~age.stra+as.factor(
grade)
          +strata(est.stra), data=gbcs)
tpc=survfit(surv_fit,
newdata=data.frame(grade=2,age.stra=3))
plot(tpc, ylim=c(0.43,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 2,
   stratified by Estrogen Receptor")
abline(0.433,0)
abline(0.642,0)
abline(0.697,0)
text(1200,.63,"less than 20", col="black")
text(2300,.63,"21 to 90", col="red")
text(2100,.82,"more than 90", col="blue")

tpd=survfit(surv_fit,
newdata=data.frame(grade=3,age.stra=3))
plot(tpd, ylim=c(0.26,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate people over 55 in tumor grade 3,
   stratified by Estrogen Receptor")
abline(0.264,0)
abline(0.492,0)
abline(0.562,0)
text(1500,.48,"less than 20", col="black")
text(1500,.62,"21 to 90", col="red")
text(1500,.82,"more than 90", col="blue")

# For Recurrence Part
rec_fit<-coxph(Surv(rectime,censrec)~age.stra+as.factor(grad
e)
          +strata(pro.stra), data=gbcs)
```

```
tpa=survfit(rec_fit,
newdata=data.frame(grade=2,age.stra=3))
plot(tpa, ylim=c(0.14,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 2,
   stratified by Progesterone Receptor")
abline(0.137,0)
abline(0.378,0)
abline(0.514,0)
text(1500,.37,"less than 20", col="black")
text(1500,.52,"21 to 90", col="red")
text(1500,.73,"more than 90", col="blue")

tpb=survfit(rec_fit,
newdata=data.frame(grade=3,age.stra=3))
plot(tpb, ylim=c(0.13,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 3,
   stratified by Progesterone Receptor")
abline(0.128,0)
abline(0.365,0)
abline(0.499,0)
text(1500,.34,"less than 20", col="black")
text(1500,.5,"21 to 90", col="red")
text(1500,.73,"more than 90", col="blue")

#for Estrogen
#par(mfrow=c(1,2))
rec_fit<-coxph(Surv(rectime,censrec)~age.stra+as.factor(grad
e)+strata(est.stra), data=gbcs)
tpc=survfit(rec_fit, newdata=data.frame(grade=2,age.stra=3))
plot(tpc, ylim=c(0.26,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 2,
   stratified by Estrogen Receptor")
abline(0.266,0)
abline(0.393,0)
abline(0.347,0)
text(1200,.45,"less than 20", col="black")
text(1400,.61,"21 to 90", col="red")
text(1400,.72,"more than 90", col="blue")

tpd=survfit(rec_fit,
newdata=data.frame(grade=3,age.stra=3))
plot(tpd, ylim=c(0.2,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for people over 55 in tumor grade 3,
   stratified by Estrogen Receptor")
abline(0.203,0)
abline(0.324,0)
abline(0.276,0)
text(1200,.4,"less than 20", col="black")
text(1500,.52,"21 to 90", col="red")
text(1500,.62,"more than 90", col="blue")
```

```
## 3 Check the influence of hormone therapy
##check Progesterone
par(mfrow=c(2,2))
surv_fit<-coxph(Surv(survtime,censdead)~age.stra+size.stra+
pro.stra+strata(hormone), data=gbcs)
tpe=survfit(surv_fit,
newdata=data.frame(age.stra=3,size.stra=2,pro.stra=1))
plot(tpe, ylim=c(0.354,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women over 55 in tumor size 2
   and lower receptors, stratified by hormone")
abline(0.358,0)
text(1250,.54,"Treated", col="black")
text(1700,.7,"Nontreated", col="red")

tpf=survfit(surv_fit,
newdata=data.frame(age.stra=3,size.stra=2,pro.stra=2))
plot(tpf, ylim=c(0.65,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women over 55 in tumor size 2
   and middle receptors, stratified by hormone")
abline(0.648,0)
text(1250,.76,"Treated", col="black")
text(1700,.85,"Nontreated", col="red")
```

```
tpg=survfit(surv_fit,
newdata=data.frame(age.stra=3,size.stra=3,pro.stra=1))
plot(tpg, ylim=c(0.21,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
    lty = c(1, 7),
    main="Survival Rate for women over 55 in tumor size 3
    and lower receptors, stratified by hormone")
abline(0.218,0)
text(1100,.47,"Treated", col="black")
text(1710,.6,"Nontreated", col="red")

tph=survfit(surv_fit,
newdata=data.frame(age.stra=1,size.stra=2,pro.stra=1))
plot(tph, ylim=c(0.38,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
    lty = c(1, 7),
    main="Survival Rate for women under 45 in tumor size 2
    and lower receptors, stratified by hormone")
abline(0.381,0)
text(1250,.48,"Treated", col="black")
text(2190,.62,"Nontreated", col="red")

# check Estrogen
#par(mfrow=c(1,2))
surv_fit<-coxph(Surv(survtime,censdead)~age.stra+size.stra+
est.stra+strata(hormone), data=gbcs)
tpe1=survfit(surv_fit,
newdata=data.frame(age.stra=2,size.stra=3,est.stra=1))
plot(tpe1, ylim=c(0.45,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
    lty = c(1, 7),
    main="Survival Rate for women over 55 in tumor size 2
    and lower receptors, stratified by hormone")
abline(0.46,0)
text(1550,.57,"Treated", col="black")
text(1650,.73,"Nontreated", col="red")

tpf1=survfit(surv_fit,
newdata=data.frame(age.stra=2,size.stra=3,est.stra=2))
plot(tpf1, ylim=c(0.58,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
    lty = c(1, 7),
    main="Survival Rate for women over 55 in tumor size 2
    and middle receptors, stratified by hormone")
abline(0.59,0)
text(1550,.7,"Treated", col="black")
text(1750,.82,"Nontreated", col="red")

tpg1=survfit(surv_fit,
newdata=data.frame(age.stra=3,size.stra=3,est.stra=1))
plot(tpg1, ylim=c(0.28,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
    lty = c(1, 7),
    main="Survival Rate for women over 55 in tumor size 2
    and lower receptors, stratified by hormone")
abline(0.28,0)
text(1550,.38,"Treated", col="black")
text(1750,.62,"Nontreated", col="red")

tph1=survfit(surv_fit,
newdata=data.frame(age.stra=3,size.stra=1,est.stra=1))
plot(tph1, ylim=c(0.35,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
    lty = c(1, 7),
    main="Survival Rate for women under 45 in tumor size 3
    and lower receptors, stratified by hormone")
abline(0.353,0)
text(1550,.46,"Treated", col="black")
text(1750,.67,"Nontreated", col="red")

##For tumor grade
##check Progesterone
par(mfrow=c(2,2))
surv_fit<-coxph(Surv(survtime,censdead)~age.stra+as.factor(
grade)+pro.stra+strata(hormone), data=gbcs)
tpe=survfit(surv_fit,
newdata=data.frame(age.stra=2,grade=3,pro.stra=1))
plot(tpe, ylim=c(0.37,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
    lty = c(1, 7),
    main="Survival Rate for women over 55 in tumor grade 2
    and lower receptors, stratified by hormone")
abline(0.368,0)
text(1250,.54,"Treated", col="black")
text(1700,.7,"Nontreated", col="red")
```

```
tpf=survfit(surv_fit,
newdata=data.frame(age.stra=2,grade=3,pro.stra=2))
plot(tpf, ylim=c(0.63,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women over 55 in tumor grade 2
   and middle receptors, stratified by hormone")
abline(0.63,0)
text(1250,.76,"Treated", col="black")
text(1700,.85,"Nontreated", col="red")

tpg=survfit(surv_fit,
newdata=data.frame(age.stra=3,grade=3,pro.stra=1))
plot(tpg, ylim=c(0.27,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women over 55 in tumor grade 3
   and lower receptors, stratified by hormone")
abline(0.275,0)
text(1250,.47,"Treated", col="black")
text(1710,.6,"Nontreated", col="red")

tph=survfit(surv_fit,
newdata=data.frame(age.stra=1,grade=2,pro.stra=1))
plot(tph, ylim=c(0.32,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women under 45 in tumor grade 2
   and lower receptors, stratified by hormone")
abline(0.315,0)
text(1250,.48,"Treated", col="black")
text(1730,.62,"Nontreated", col="red")

# check Estrogen
surv_fit<-coxph(Surv(survtime,censdead)~as.factor(grade)+ag
e.stra+est.stra+strata(hormone), data=gbcs)
tpe1=survfit(surv_fit,
newdata=data.frame(age.stra=2,grade=3,est.stra=1))
plot(tpe1, ylim=c(0.45,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women over 55 in tumor grade 2
   and lower receptors, stratified by hormone")
abline(0.46,0)
text(1550,.57,"Treated", col="black")
text(1650,.73,"Nontreated", col="red")

tpf1=survfit(surv_fit,
newdata=data.frame(age.stra=2,grade=3,est.stra=2))
plot(tpf1, ylim=c(0.58,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women over 55 in tumor grade 2
   and middle receptors, stratified by hormone")
abline(0.59,0)
text(1550,.7,"Treated", col="black")
text(1750,.82,"Nontreated", col="red")

tpg1=survfit(surv_fit,
newdata=data.frame(age.stra=3,grade=3,est.stra=1))
plot(tpg1, ylim=c(0.28,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women over 55 in tumor grade 2
   and lower receptors, stratified by hormone")
abline(0.28,0)
text(1550,.38,"Treated", col="black")
text(1750,.62,"Nontreated", col="red")

tph1=survfit(surv_fit,
newdata=data.frame(age.stra=3,grade=1,est.stra=1))
plot(tph1, ylim=c(0.35,1),

col=c("black","red","blue","green","yellow","cyan","brown"),
   lty = c(1, 7),
   main="Survival Rate for women under 45 in tumor grade 3
   and lower receptors, stratified by hormone")
abline(0.353,0)
text(1550,.46,"Treated", col="black")
text(1750,.67,"Nontreated", col="red")
```