Survival Analysis
Module III: Deep Dive

Dr. Mark Williamson

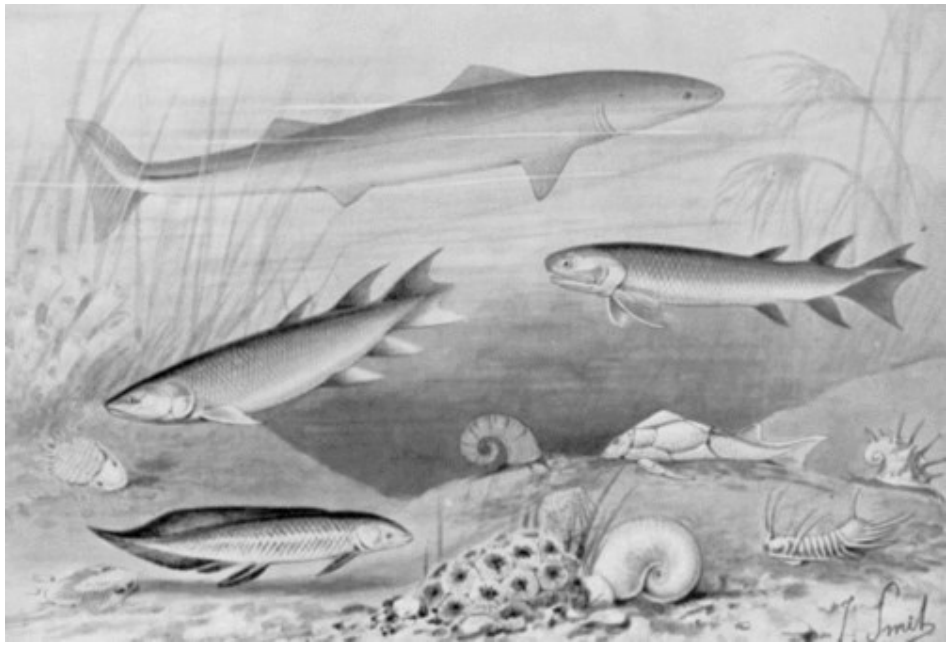
DaCCoTA

University of North Dakota

Introduction

- Last time, we wandered in the trees of survival analysis and examined how to determine survival times of a group or across groups

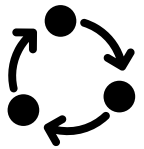
- Today, we'll swim over to examine the effects of factors on survival times via tree and regression models
- We will also look at detailed examples in R and SAS



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Reviewing the Basics

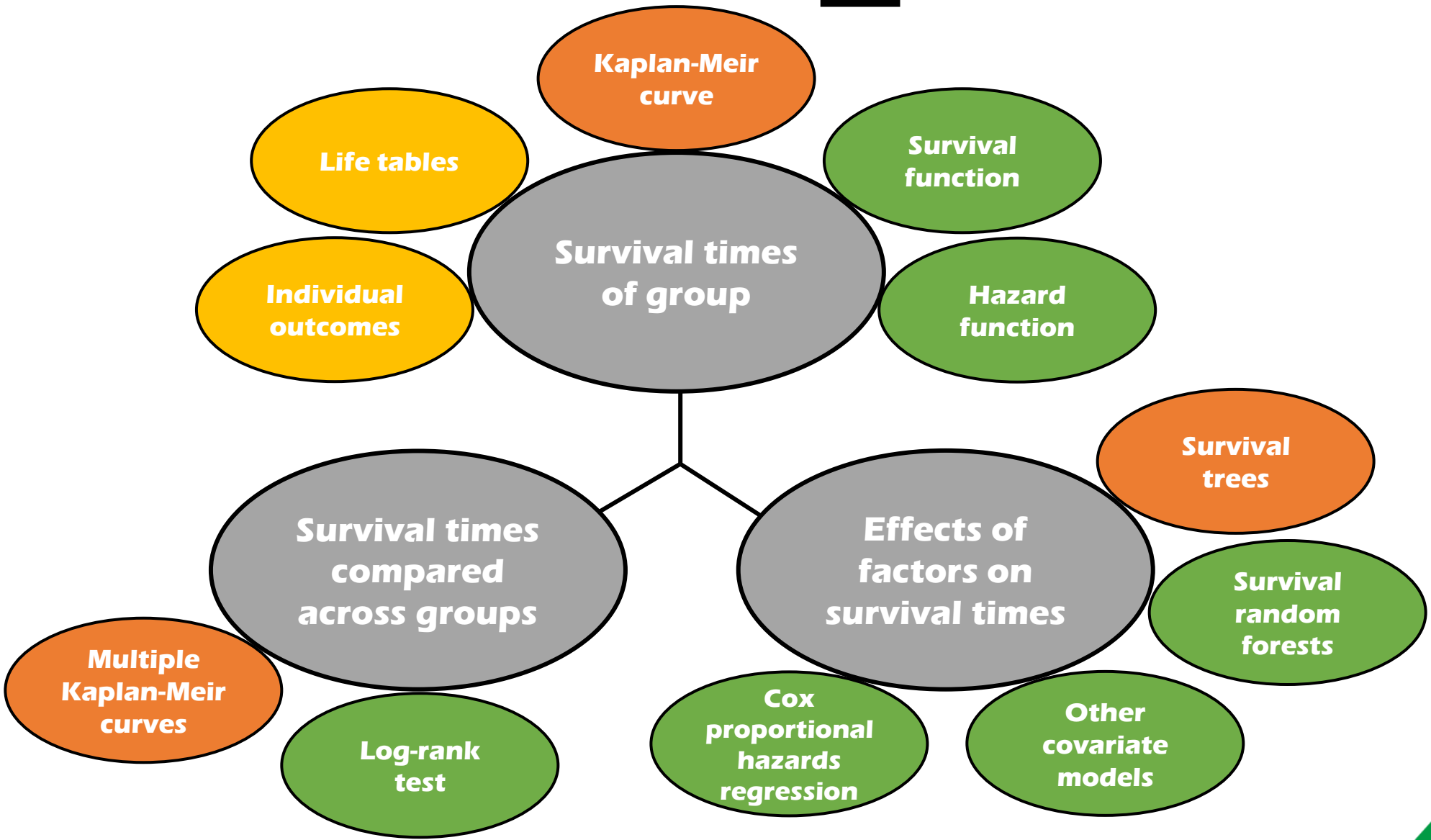


- Survival analysis uses time-to-event data (correct format needed)
- Almost always involves censoring
- Common calculations are survival and hazard function
- Kaplan-Meier curves are non-parametric
- Parametric curves also available with more math
- The curves of two groups can be compared with log-rank test
- Extending beyond two groups or involving co-variates requires tree-based methods or Cox regression

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

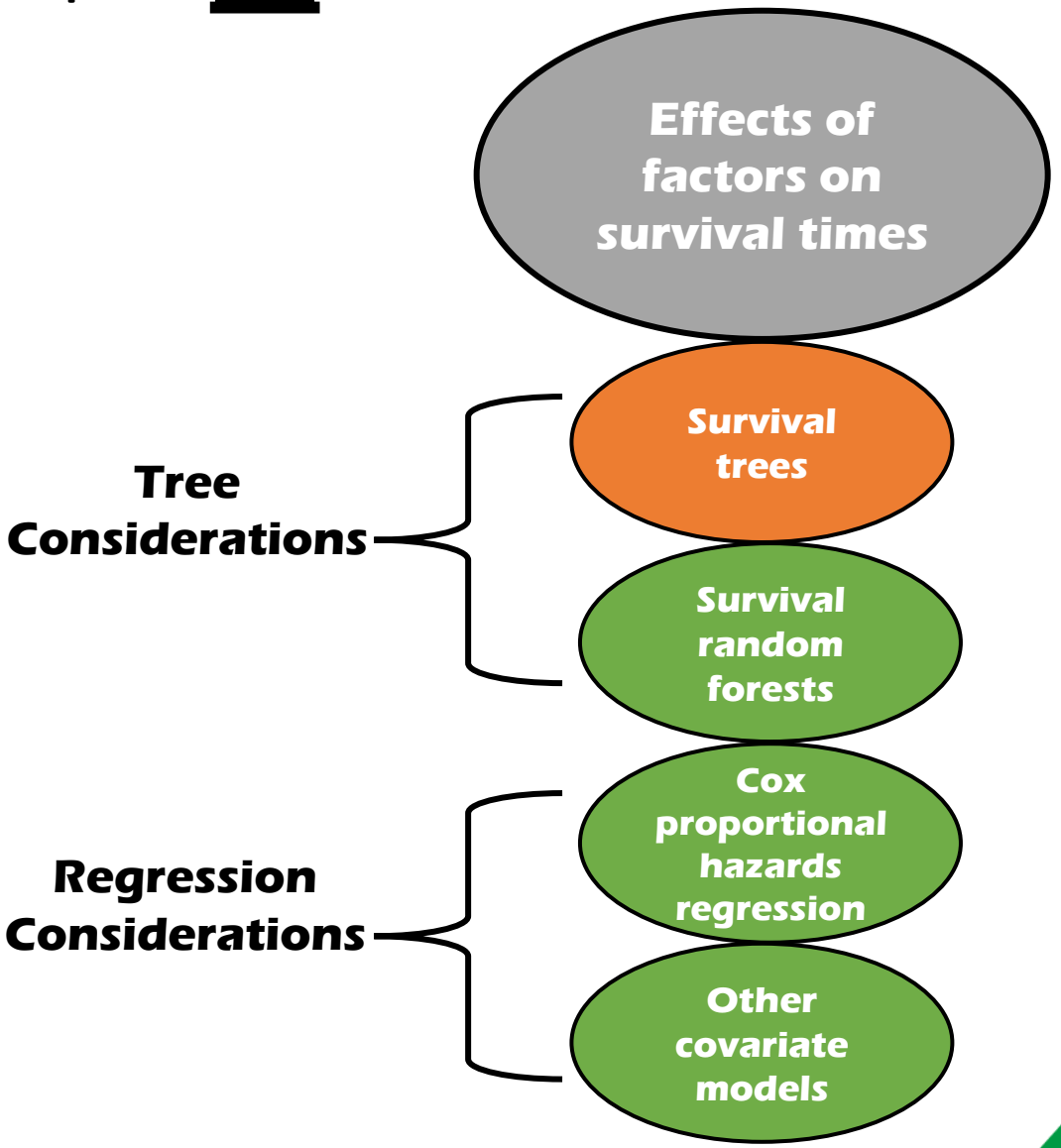
Scope



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Scope



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Regression Considerations

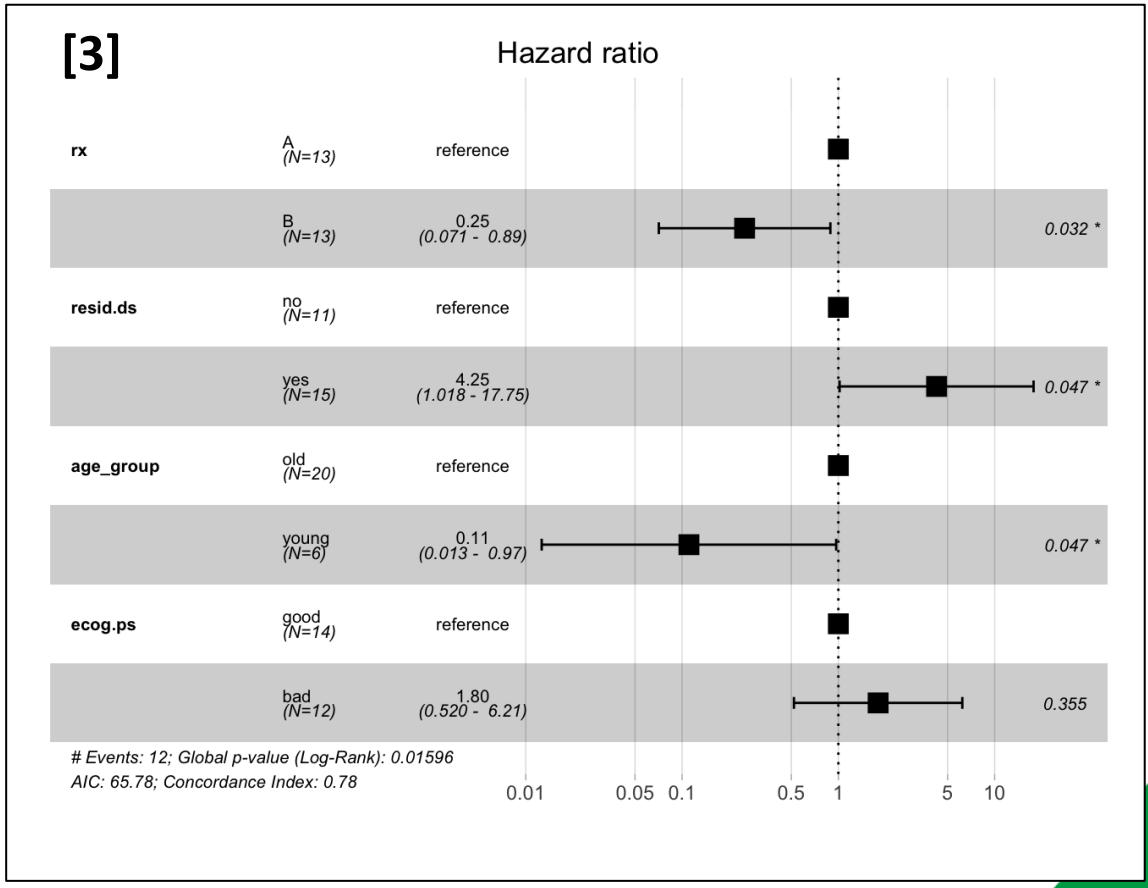
- Cox Proportional Hazards regression, aka Cox regression, is used to predicted the probability of an event (death/otherwise) across time for given values of predictor variables [1]
- Extension of basic survival analysis; analogous of going from simple to multiple regression [2]
- Better yet, think of logistic regression and odds ratios
- Relating several risk factors to survival time by measuring hazard rate [2]
- Cox model is semi-parametric
- Parametric regression models are also available

Risk Factor [2]	Parameter Estimate	P-value
Age (years)	0.11149	0.0001
Male sex	0.67958	0.0001

Risk Factor	Parameter Estimate	P-value	Hazard Radio (CI)
Age (years)	0.11691	0.0001	1.124 (1.111-1.138)
Male sex	0.40359	0.0002	1.497 (1.215-1.845)
Systolic Blood Pressure	0.01645	0.0001	1.017 (1.012-1.021)
Current Smoker	0.76798	0.0001	2.155 (1.758-2.643)
Total Serum Cholesterol	-0.00209	0.0963	0.998 (0.995-2.643)
Diabetes	-0.02366	0.1585	0.816 (0.615-1.083)

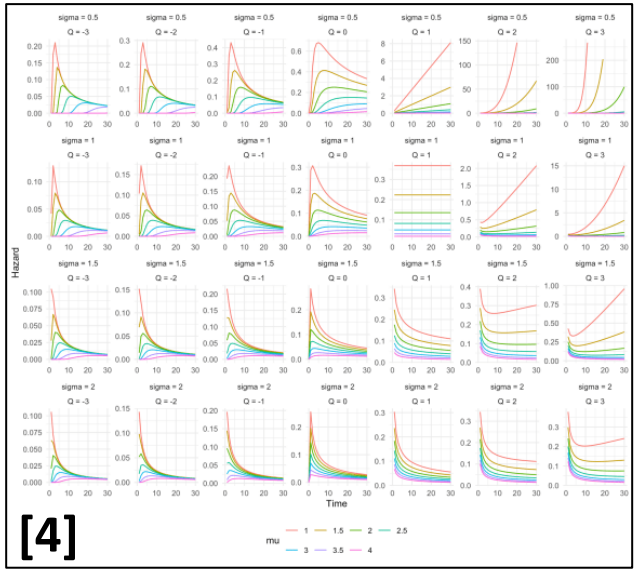
Regression Considerations

- Cox Proportional Hazards regression, aka Cox regression, is used to predicted the probability of an event (death/otherwise) across time for given values of predictor variables [1]
- Extension of basic survival analysis; analogous of going from simple to multiple regression [2]
- Better yet, think of logistic regression and odds ratios
- Relating several risk factors to survival time by measuring hazard rate [2]
- Cox model is semi-parametric
- Parametric regression models are also available

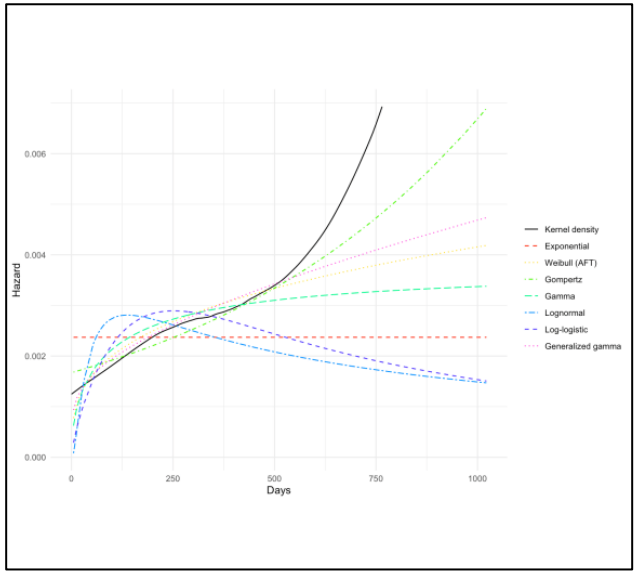


Regression Considerations 2

- Cox models do not assume any particular baseline survival distribution [4]
- Useful but have limitations (ex. poor extrapolation)
- Parametric survival models can do the trick
- Models the data in more detail by using maximum likelihood with an appropriate distribution [5]
- Links survival time or hazard of an individual to covariates using a specified probability distribution [6]
- Distributions include Weibull, exponential, log-normal, etc. [7, 8]
- Different distributions fit different hazards data
- Used for survival, hazard, and proportional hazards



[4]



Regression Considerations 3

[9] Type of Model	Baseline Hazard (Time + event)	Covariates
Non-parametric	No distribution assumed	No distribution assumed
Semi-parametric	No distribution assumed	Some distribution assumed
Parametric	Some distribution assumed	Some distribution assumed

Exponential

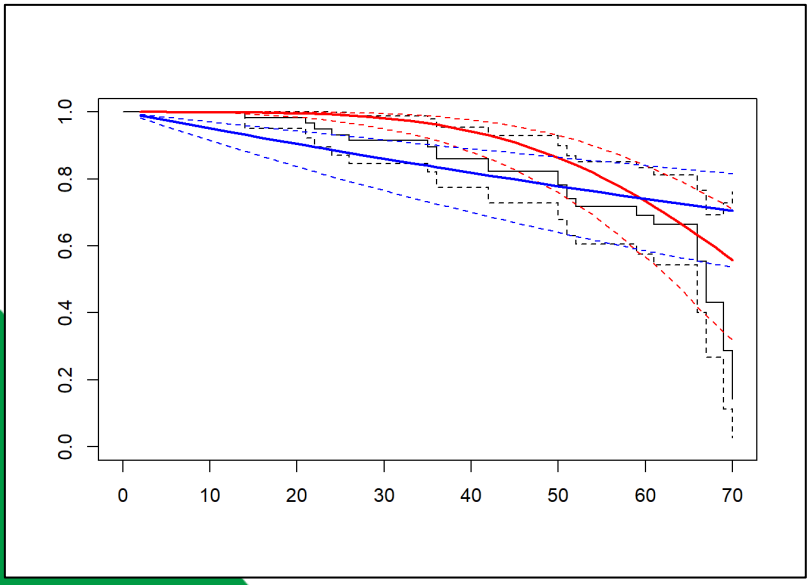
Analysis of Maximum Likelihood Parameter Estimates

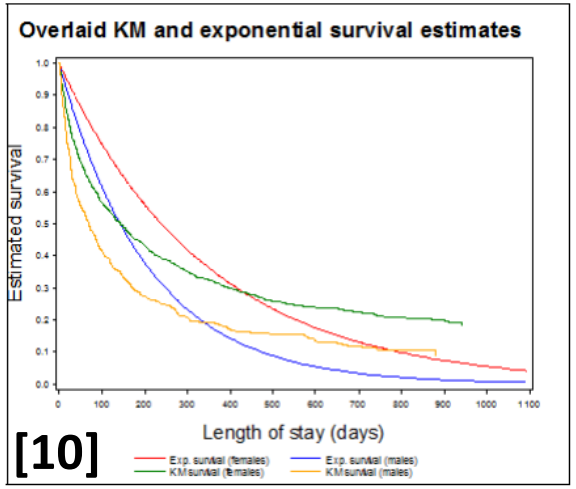
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.8421	0.0333	5.7769	5.9074	30785.8	<.0001
gender	1	-0.5162	0.0619	-0.6375	-0.3948	69.51	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		
Weibull Shape	0	1.0000	0.0000	1.0000	1.0000		

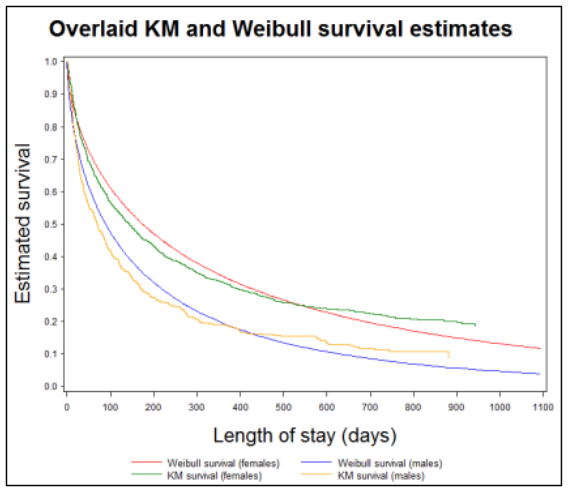
Weibull

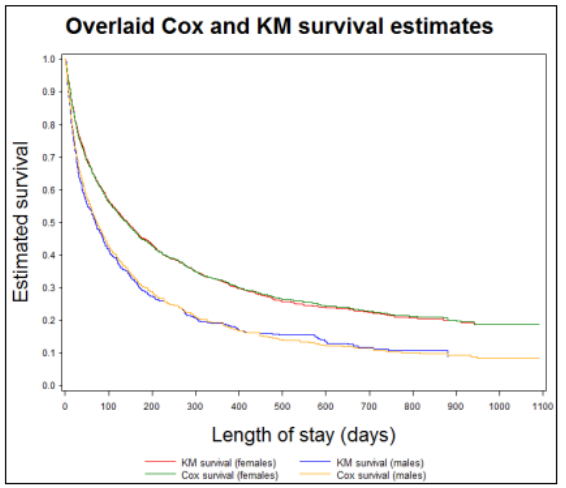
Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.7564	0.0542	5.6502	5.8627	11280.0	<.0001
gender	1	-0.6735	0.1011	-0.8716	-0.4754	44.40	<.0001
Scale	1	1.6275	0.0378	1.5551	1.7032		
Weibull Shape	1	0.6144	0.0143	0.5871	0.6430		







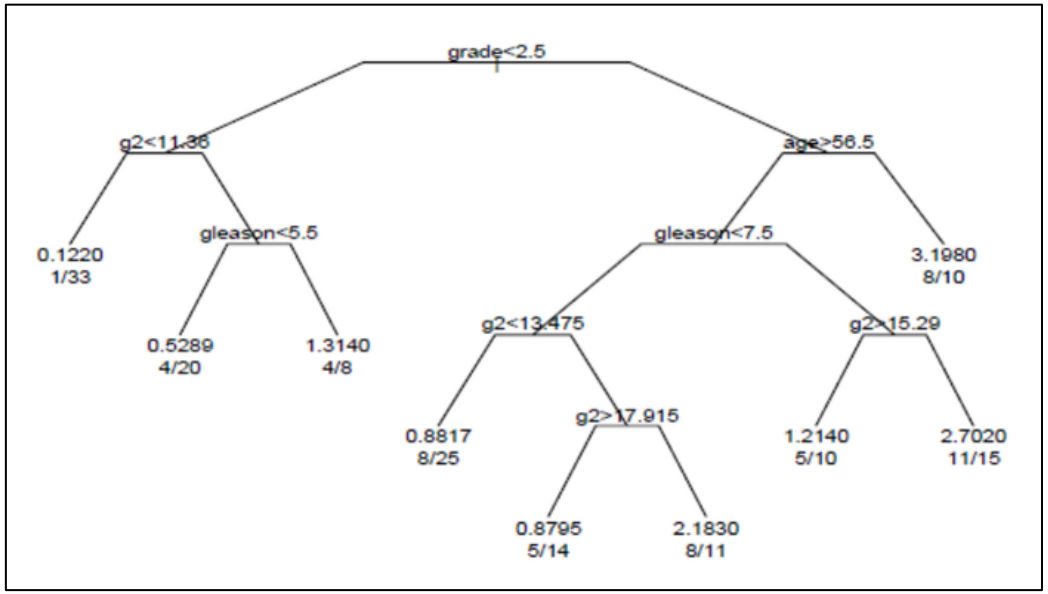


DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Tree Considerations

- Survival Tree Analysis
  - Alternative method to Cox regression, which tries to separate events (0/1) via partitioning [11]
  - Each branch indicates a split via a value of the variable
  - Ends are called terminal nodes, and they indicate the number of subjects with an event
- Survival Random Forest
  - Extension of survival tree analysis [12]
  - Instead of building one tree, many trees are built and then averaged
  - Prediction errors are estimated by resampling (bootstrap)
  - Survival and hazard curves can be generated



Other Considerations

What if you have multiple events?

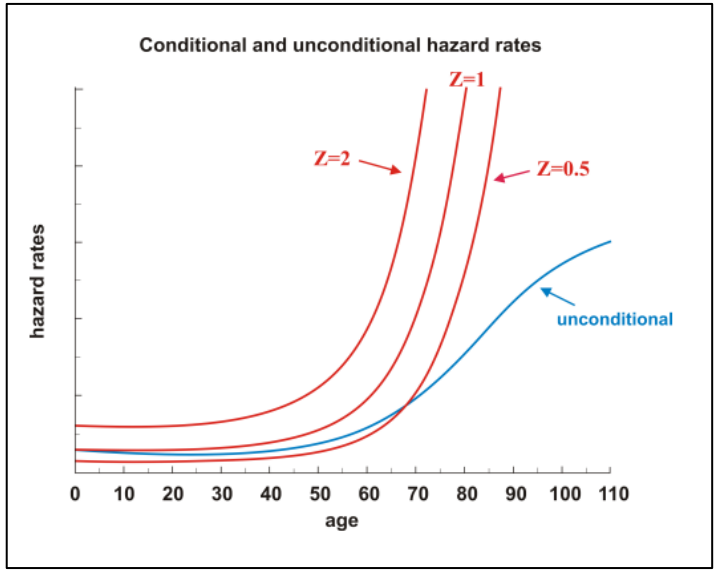
- Ex. Multiple tumor sites or heterogenous populations
- **Frailty models**
- Account for correlation within a group by introducing 'frailty term' as random effect [13, 14]
- Idea is to consider event (ex. tumor) as locus and the person/family/liter as the random effect [15]
- More broadly for accounting for the fact that individuals do not have the same underlying risk [16]

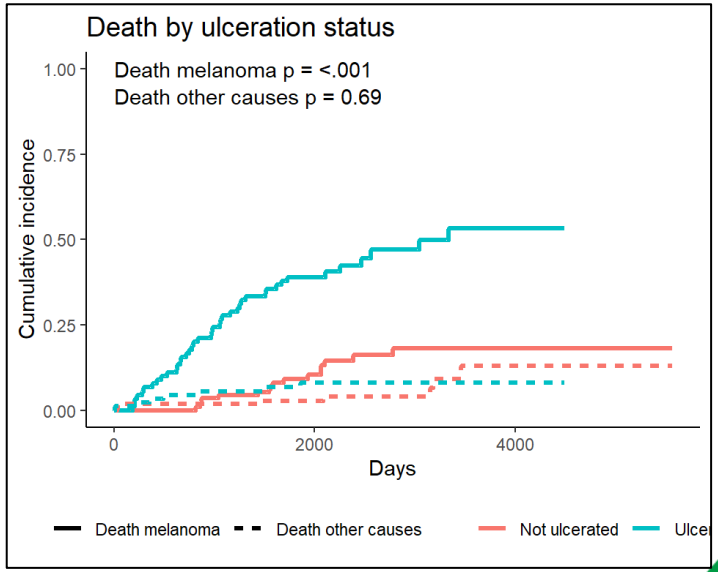
What if there are multiple ways to die?

- Ex. death from breast cancer or stroke
- **Competing risks**
- Partitions events that occur in model into discrete competing events [14]
- Cumulative Index function most used [17]

What if time is discrete rather than continuous?

- Ex. time until graduation measured in semesters
- **Logistic regression**
- Tricky bit is setting up data so that censoring is incorporated [14]
- Helpful introduction here [18]





Other Considerations 2

What if you are interested in a covariate measure **after** follow-up?

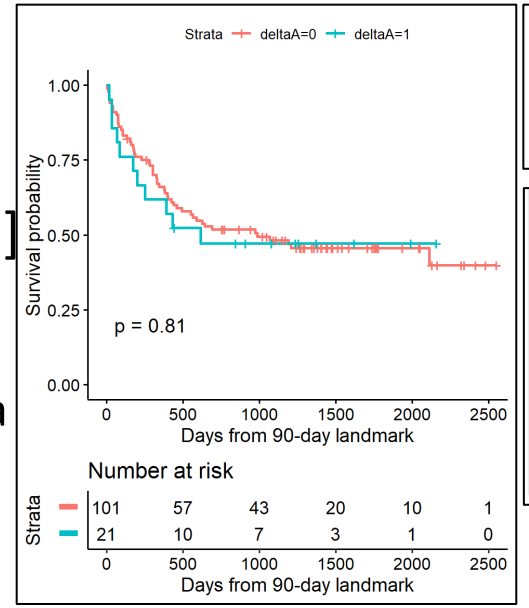
- Ex. tumor response to treatment
- **Landmark** or **time-dependent covariate** analysis [19]
  - **Landmark** -> select fixed time after baseline, subset population, Cox regression as normal
  - **Time-dependent** -> more work involved setting up data with multiple time intervals for each patient to account for different levels of covariate across time [20]

What about survival estimates for those who have already survived a given time or event?

- **Conditional survival**
- Different curves for different conditions [19]

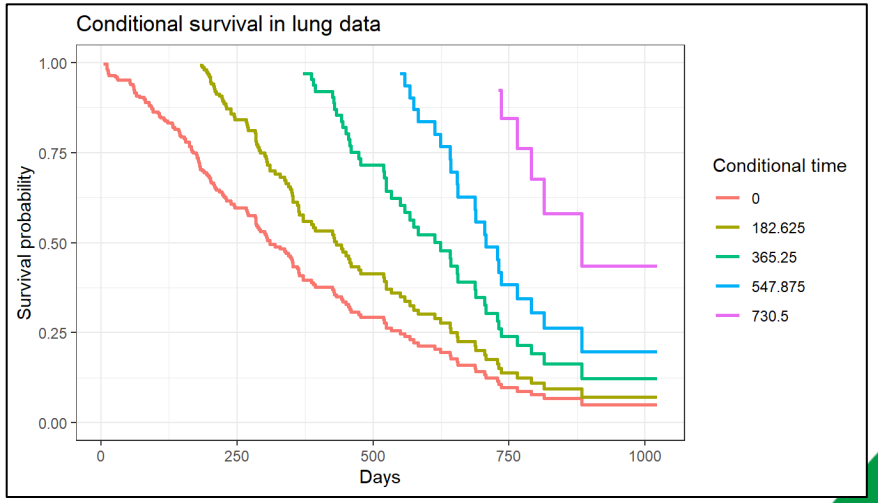
What if you're an obligate Bayesian?

- **Bayesian** methods available too [21]



	subject	time1	time2	death	creatinine
1	5	0	90	0	0.9
2	5	90	120	0	1.5
3	5	120	185	1	1.2

##	my_id	T1	delta1	id	tstart	tstop	death	agvhd
## 1	1	2081	0	1	0	67	0	0
## 2	1	2081	0	1	67	2081	0	1
## 3	2	1602	0	2	0	1602	0	0
## 4	3	1496	0	3	0	1496	0	0
## 5	4	1462	0	4	0	70	0	0
## 6	4	1462	0	4	70	1462	0	1
## 7	5	1433	0	5	0	1433	0	0



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Assessment 1

qualtrics XM

UNIVERSITY OF NORTH DAKOTA UND

https://und.qualtrics.com/jfe/form/SV_0xIqRRaWWG99Fga

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1.1

Survival Analysis in R

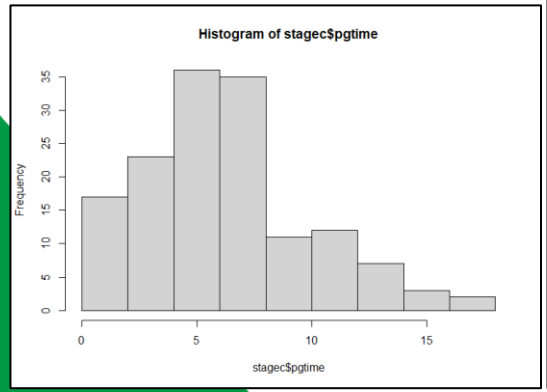
Prostate Cancer: Cox regression with single variable

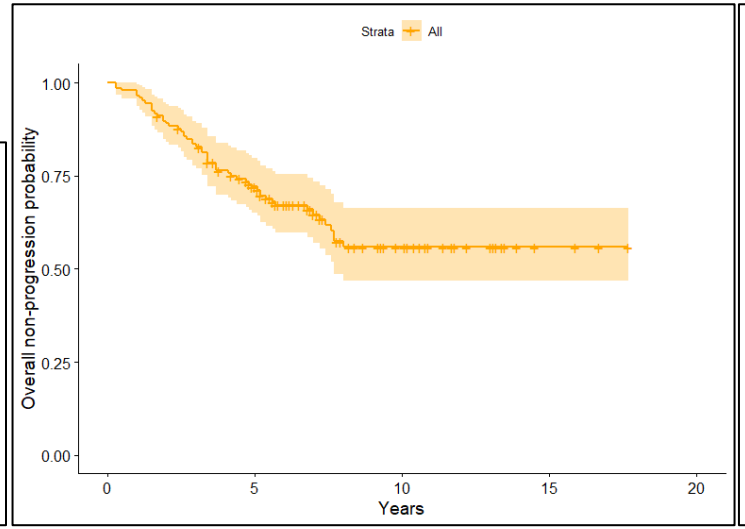
```
library(ggplot2)
library(survival)
library(survminer)
library(broom)
library(knitr)
library(gtsummary)
library(muhaz)
library(flexsurv)
library(data.table)
library(rpart)
library(randomForestSRC)
library(MASS)
```

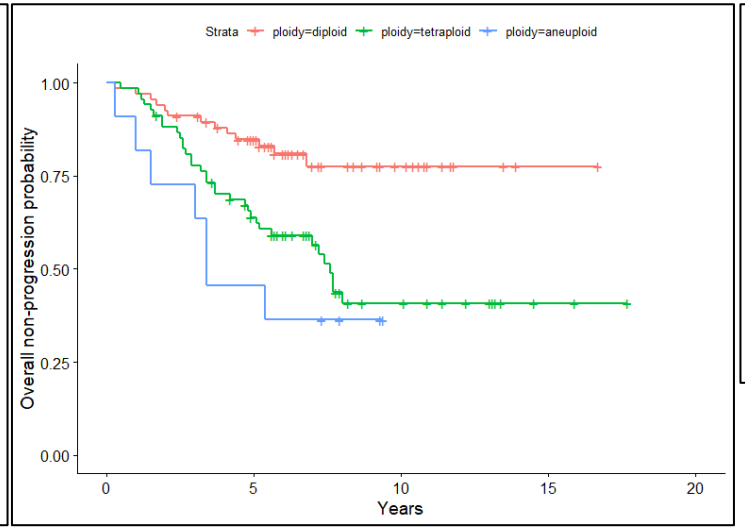
```
head(stagec)
hist(stagec$pgtime) #years
table(stagec$pgstat)
fit1 <- ggsurvplot(
  fit = survfit(Surv(pgtime, pgstat)~1, data=stagec),
  xlab="Years",
  ylab="Overall non-progression probability",
  palette = "orange")
fit1
```

0	1
92	54

```
mod1 <-coxph(Surv(pgtime,pgstat)~ploidy, data=stagec)
mod1 %>%
  gtsummary::tbl_regression(exp=TRUE)
fit2 <- ggsurvplot(
  fit = survfit(Surv(pgtime, pgstat)~ploidy, data=stagec),
  xlab="Years",
  ylab="Overall non-progression probability")
fit2
```







Characteristic	HR[1]	95% CI[1]	p-value
ploidy			
diploid	—	—	
tetraploid	2.84	1.50, 5.39	0.001
aneuploid	4.34	1.73, 10.9	0.002

[1] HR = Hazard Ratio, CI = Confidence Interval

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1.2

Prostate Cancer: Cox regression with multiple variables

```
mod2.1 <-coxph(Surv(pgtime, pgstat)~ploidy
               + age + factor(eet)
               + g2 + factor(grade) + factor(gleason),
               data=stagec)
mod2.1 %>%
 gtsummary::tbl_regression(exp=TRUE)

mod2.2 <-coxph(Surv(pgtime, pgstat)~ploidy
               + age + factor(eet)
               + g2 + grade + gleason,
               data=stagec)
mod2.2 %>%
 gtsummary::tbl_regression(exp=TRUE)

mod2.3 <-coxph(Surv(pgtime, pgstat)~ploidy
               + g2,
               data=stagec)
mod2.3 %>%
 gtsummary::tbl_regression(exp=TRUE)
```

Characteristic	HR[1]	95% CI[1]	p-value
ploidy			
diploid	—	—	
tetraploid	3.23	1.78, 5.88	<0.001
aneuploid	5.39	1.93, 15.0	0.001
age	0.99	0.94, 1.05	0.8
factor(eet)			
1	—	—	
2	1.22	0.64, 2.33	0.5
g2	0.95	0.92, 0.99	0.006
factor(grade)			
1	—	—	
2	5,491,404	2,561,074, 11,774,560	<0.001
3	10,958,263	5,688,299, 21,110,621	<0.001
4	59,372,886	20,437,407, 172,484,680	<0.001
factor(gleason)			
3	—	—	
4	0.08	0.02, 0.31	<0.001
5	0.02	0.01, 0.06	<0.001
6	0.05	0.02, 0.09	<0.001
7	0.07	0.04, 0.13	<0.001
8	0.06	0.03, 0.13	<0.001
9	0.08	0.02, 0.25	<0.001
10	1.00	0.16, 6.20	>0.9

[1] HR = Hazard Ratio, CI = Confidence Interval

Characteristic	HR[1]	95% CI[1]	p-value
ploidy			
diploid	—	—	
tetraploid	2.69	1.17, 6.20	0.020
aneuploid	3.39	0.92, 12.5	0.067
age	0.98	0.93, 1.04	0.5
factor(eet)			
1	—	—	
2	1.09	0.53, 2.28	0.8
g2	0.95	0.90, 1.00	0.037
grade	3.96	1.69, 9.32	0.002
gleason	1.23	0.88, 1.73	0.2

[1] HR = Hazard Ratio, CI = Confidence Interval

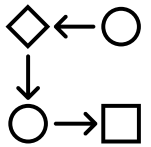
Characteristic	HR[1]	95% CI[1]	p-value
ploidy			
diploid	—	—	
tetraploid	4.20	1.90, 9.31	<0.001
aneuploid	7.17	2.31, 22.2	<0.001
g2	0.97	0.92, 1.01	0.2

[1] HR = Hazard Ratio, CI = Confidence Interval

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1.3



Prostate Cancer: Parametric models [4]

```
k_haz_est <-muhaz(stagec$pgtime, stagec$pgstat)

k_haz <-data.table(time=k_haz_est$est.grid,
                   est= k_haz_est$haz.est,
                   method="Kernel density")

dists <- c("exp", "weibull", "gompertz", "gamma",
           "lognormal", "llogis", "gengamma")

dists_long <- c("Exponential", "Weibull (AFT)",
                "Gompertz", "Gamma", "Lognormal", "Log-logistic",
                "Generalized gamma")

parametric_haz <- vector(mode = "list", length = length(dists))
```

```
for (i in 1:length(dists)){
  fit <- flexsurvreg(Surv(pgtime, pgstat) ~ 1, data = stagec, dist = dists[i])
  parametric_haz[[i]] <- summary(fit, type = "hazard",
                                 ci = FALSE, tidy = TRUE)
  parametric_haz[[i]]$method <- dists_long[i]
}

parametric_haz <- rbindlist(parametric_haz)

haz <- rbind(k_haz, parametric_haz)

haz[, method := factor(method,
                       levels = c("Kernel density",
                                  dists_long))]

n_dists <- length(dists)
```

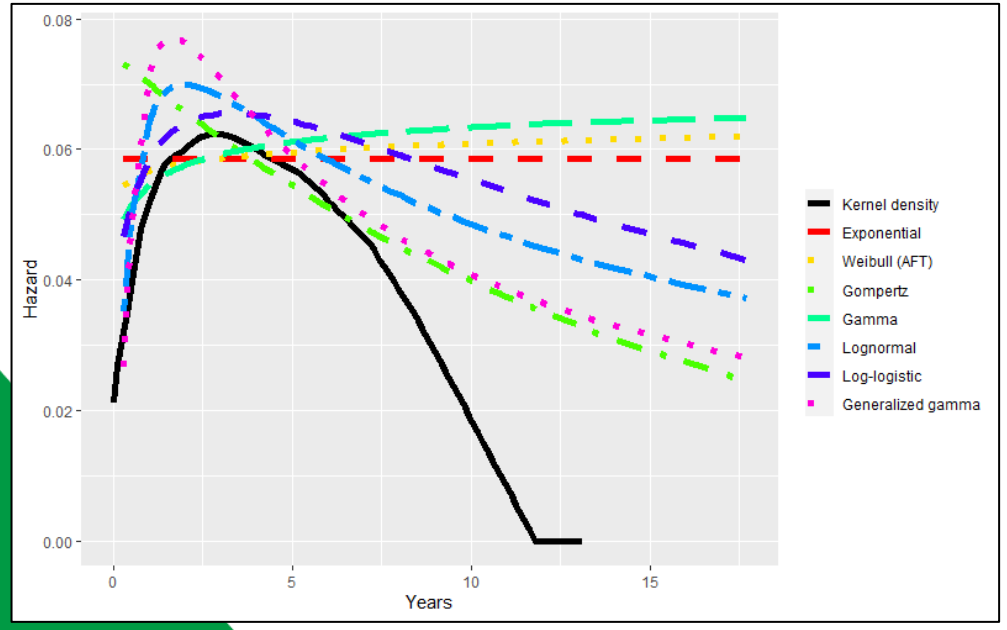
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1.3

Prostate Cancer: Parametric models

```
ggplot(haz, aes(x = time, y = est, col = method, linetype = method)) +
  geom_line(size=2) +
  xlab("Years") + ylab("Hazard") +
  scale_colour_manual(name = "",
                      values = c("black", rainbow(n_dists))) +
  scale_linetype_manual(name = "",
                        values = c(1, rep_len(2:6, n_dists)))
```




```
mod3 <-flexsurvreg(Surv(pgtime, pgstat)~ploidy,
                   data=stagec, dist='gengamma')
mod3 %>%
  gtsummary::tbl_regression(exp=TRUE)
```

Characteristic	exp(Beta)	95% CI[1]	p-value
mu	3.05	2.32, 3.78	
sigma	1.42	1.01, 2.00	
Q	-0.05	-1.02, 0.93	
ploidytetraploid	-0.85	-1.48, -0.21	0.004
ploidyaneuploid	-1.56	-2.57, -0.55	0.001

[1] CI = Confidence Interval

```
mod4 <-flexsurvreg(Surv(pgtime, pgstat)~ploidy + g2 + factor(grade),
                   data=stagec, dist='gengamma')
mod4 %>%
  gtsummary::tbl_regression(exp=TRUE)
```

Characteristic	exp(Beta)	95% CI[1]	p-value
mu	6.12	-133, 146	
sigma	1.18	1.02, 1.37	
Q	-0.56	-1.48, 0.37	
ploidytetraploid	-1.17	-1.81, -0.54	<0.001
ploidyaneuploid	-1.62	-2.75, -0.49	0.002
g2	0.07	0.03, 0.11	<0.001
factor(grade)2	-3.51	-143, 136	0.5
factor(grade)3	-4.62	-144, 135	0.5
factor(grade)4	-6.76	-146, 133	0.5

[1] CI = Confidence Interval

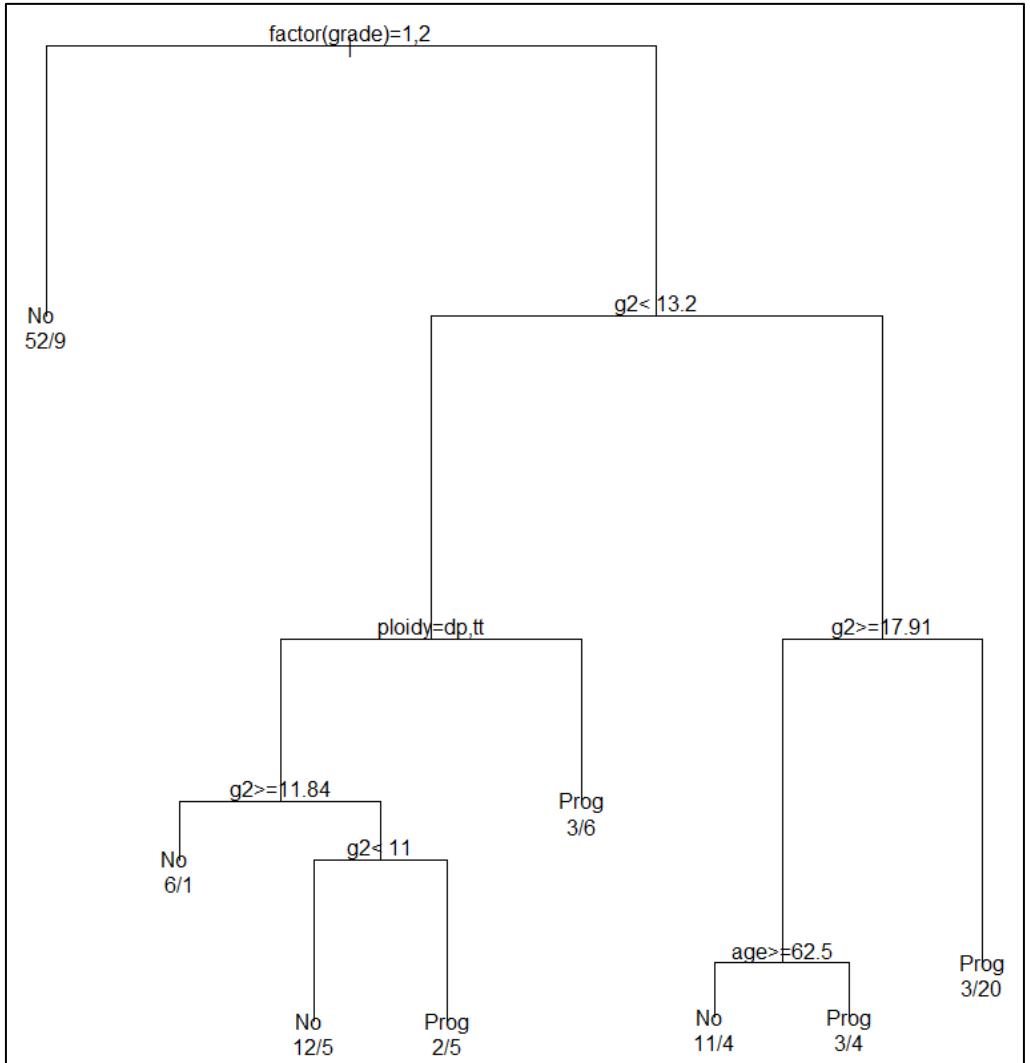
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1.4

Prostate Cancer: Survival Tree

```
stagec$progstat <- factor(stagec$pgstat, levels = 0:1,
                          labels = c("No", "Prog"))
tree1 <- rpart(progstat ~ age + factor(eet)
               + g2 + factor(grade) + factor(gleason) + ploidy,
               data = stagec, method = 'class')
par(mar=rep(0.1, 4))
plot(tree1)
text(tree1, use.n=TRUE, min=2)
par(mar=c(5.1, 4.1, 4.1, 2.1))
```



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

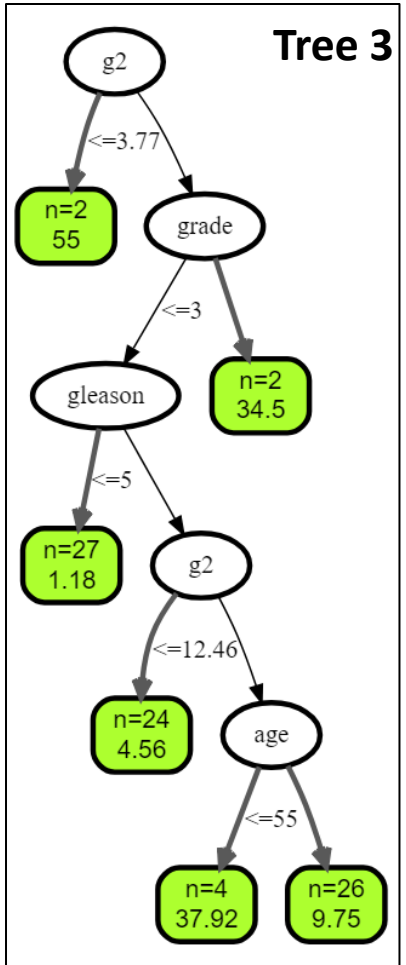
UNIVERSITY OF NORTH DAKOTA

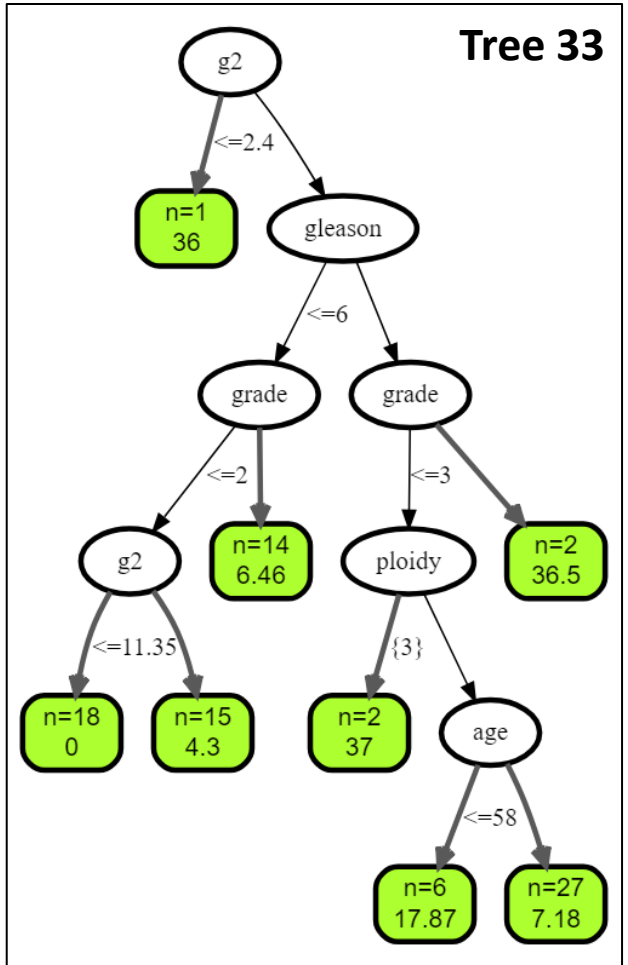
Step-by-step Example 1.5

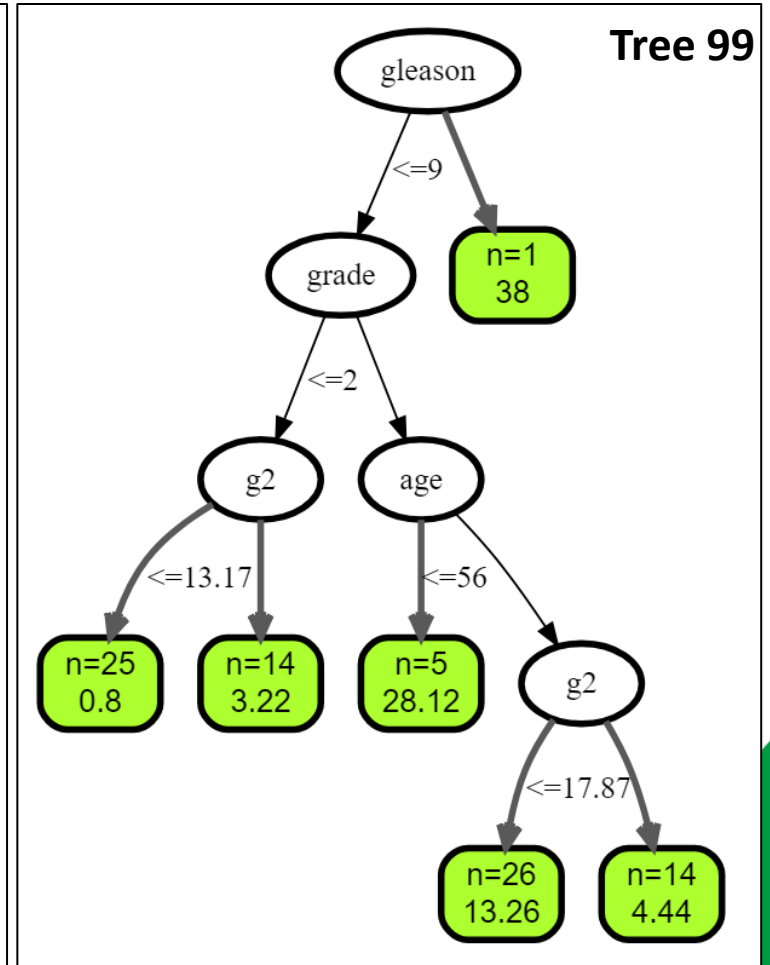
Prostate Cancer: Survival Random Forest

```
v.obj <-rfsrc(Surv(pgtime, pgstat) ~ age
        + factor(eet) + g2 + factor(grade)
        + factor(gleason) + ploidy,
        data=stagec,
        ntree=100, block.size=1)

plot(get.tree.rfsrc(v.obj, 3))
plot(get.tree.rfsrc(v.obj, 33))
plot(get.tree.rfsrc(v.obj, 99))
```







DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

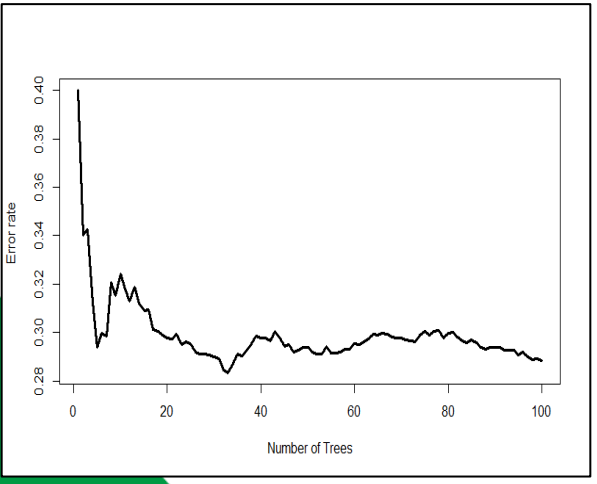
Step-by-step Example 1.5

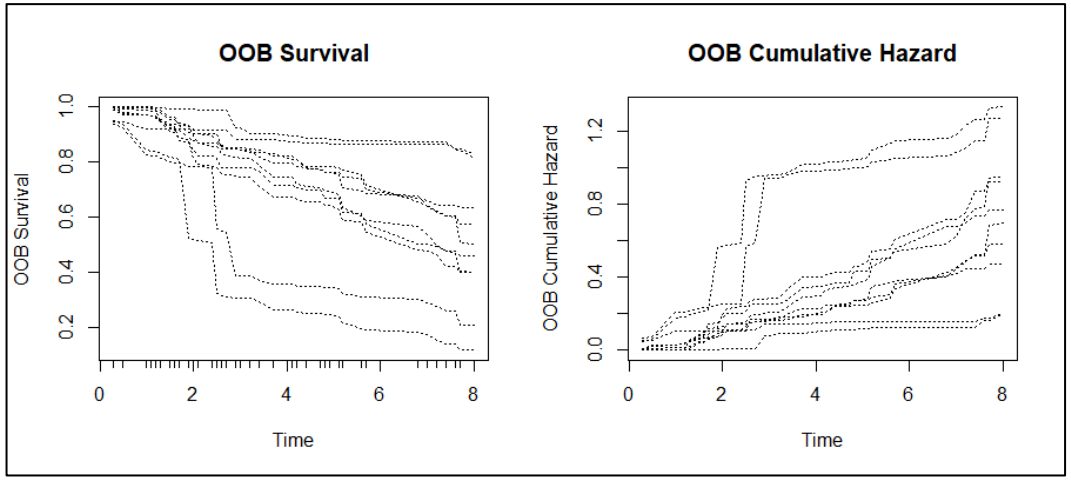
Prostate Cancer: Survival Random Forest

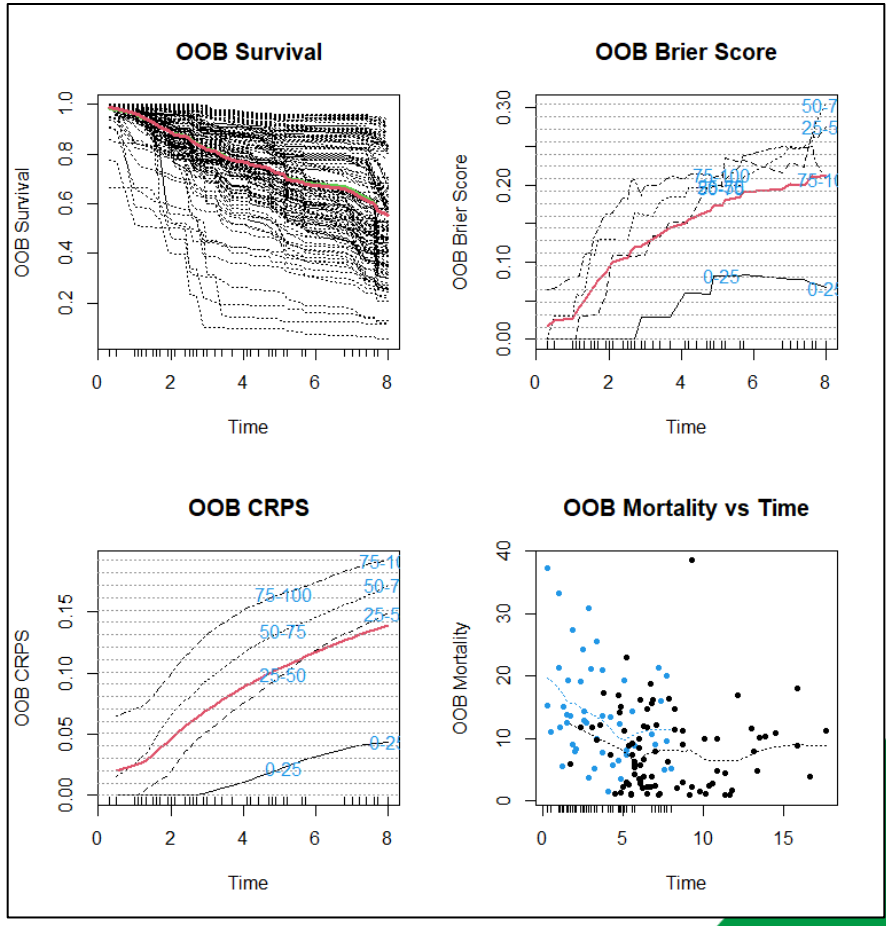
print(v.obj)
plot(v.obj)

plot.survival(v.obj, subset=1:10)
plot.survival(v.obj)

```
                         Sample size: 134
                      Number of deaths: 49
                      Number of trees: 100
            Forest terminal node size: 15
        Average no. of terminal nodes: 7.41
No. of variables tried at each split: 3
              Total no. of variables: 6
      Resampling used to grow trees: swor
   Resample size used to grow trees: 85
                           Analysis: RSF
                             Family: surv
                   Splitting rule: logrank *random*
   Number of random split points: 10
                 (OOB) Error rate: 29.20688208%
```







DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

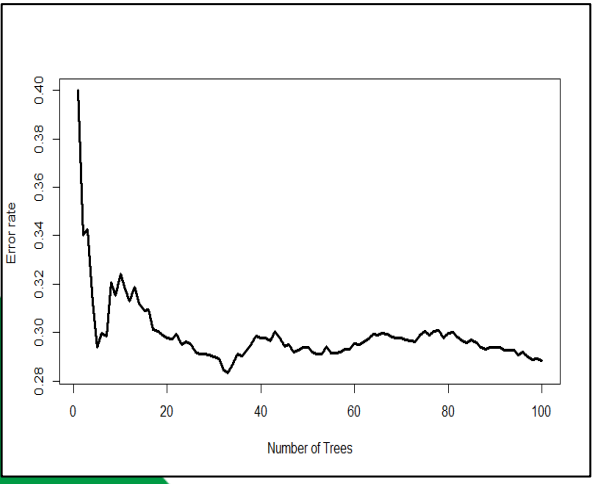
Step-by-step Example 1.5

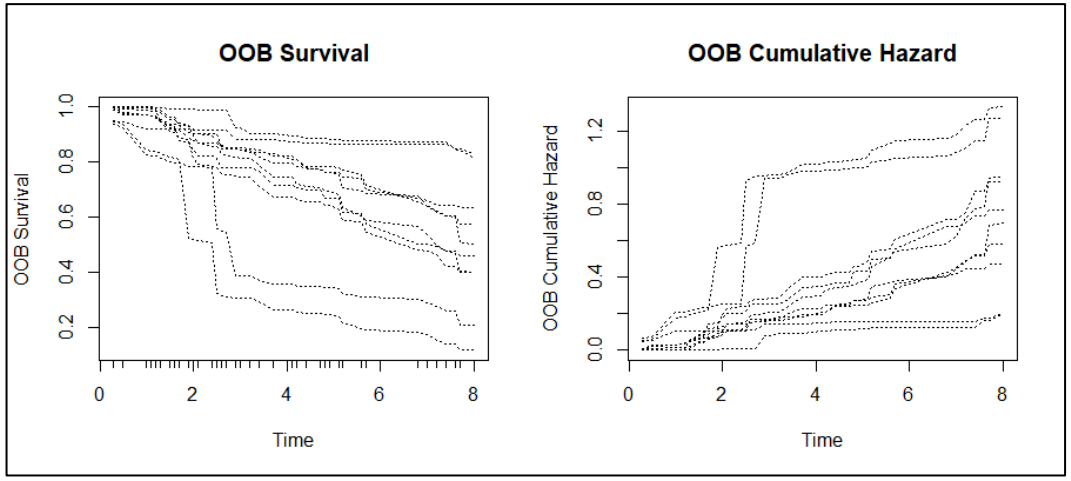
Prostate Cancer: Survival Random Forest

print(v.obj)
plot(v.obj)

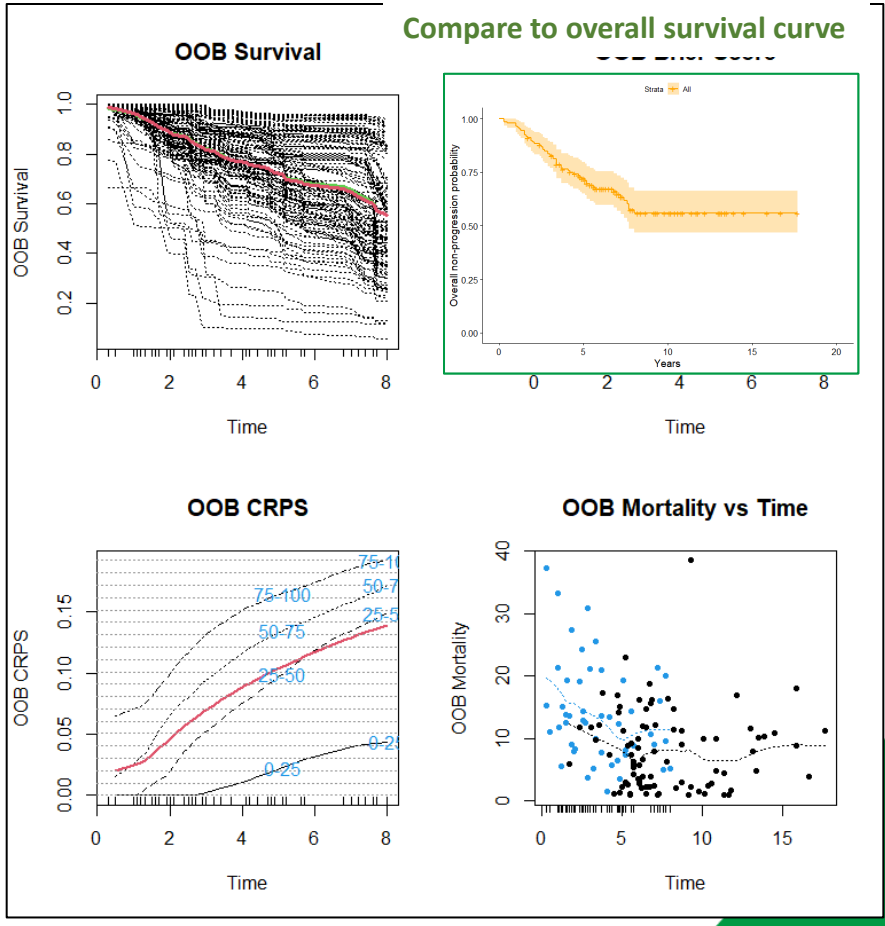
plot.survival(v.obj, subset=1:10)
plot.survival(v.obj)

```
                         Sample size: 134
                 Number of deaths: 49
                  Number of trees: 100
       Forest terminal node size: 15
   Average no. of terminal nodes: 7.41
No. of variables tried at each split: 3
         Total no. of variables: 6
   Resampling used to grow trees: swor
Resample size used to grow trees: 85
                        Analysis: RSF
                          Family: surv
                Splitting rule: logrank *random*
Number of random split points: 10
                (OOB) Error rate: 29.20688208%
```





Compare to overall survival curve



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1.6

Lung Cancer: Frailty Model [22]

```
head(lung)

 #no random effect
mod5.1 <-coxph(Surv(time, status) ~ age,
          data=lung)
mod5.1 %>%
 gtsummary::tbl_regression(exp=TRUE)

 #random institutional effect
mod5.2 <-coxph(Surv(time, status) ~ age
          + frailty(inst, df=4),
          data=lung)
mod5.2 %>%
 gtsummary::tbl_regression(exp=TRUE)
```

Characteristic	HR[1]	95% CI[1]	p-value
age	1.02	1.00, 1.04	0.042

[1] HR = Hazard Ratio, CI = Confidence Interval

Characteristic	HR[1]	95% CI[1]	p-value
age	1.02	1.00, 1.04	0.038
frailty(inst, df = 4)			0.5

[1] HR = Hazard Ratio, CI = Confidence Interval

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

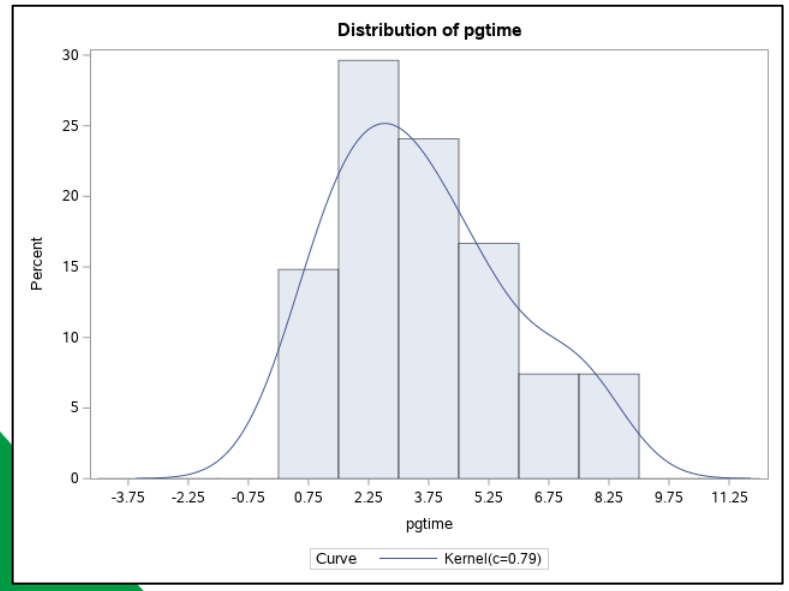
UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.1

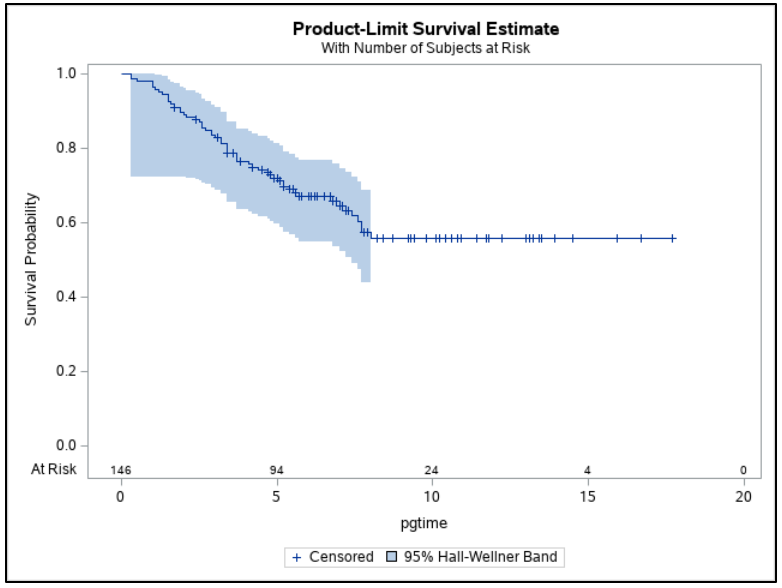
Survival Analysis in SAS

Prostate Cancer Remix: Cox regression with single variable

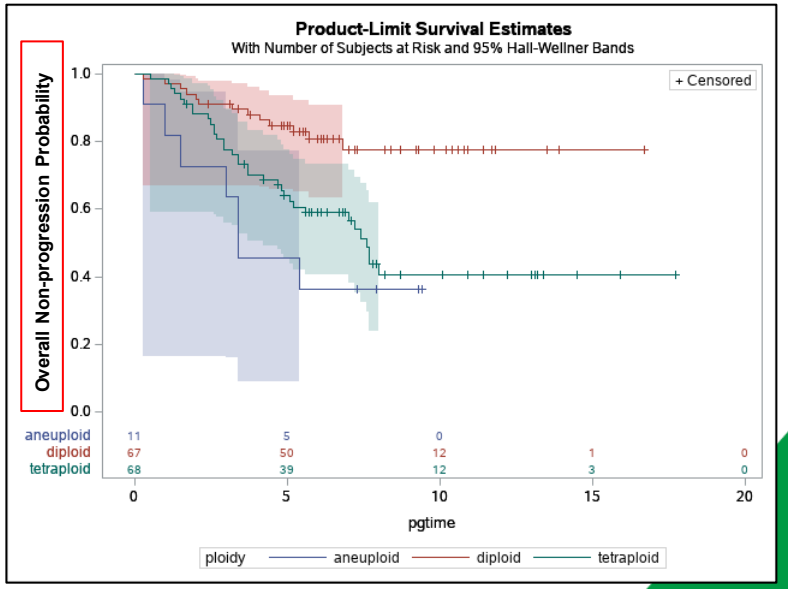
```
PROC UNIVARIATE data =stagec;
        where pgstat=1;
        var pgtime;
        histogram pgtime / kernel;
```




```
PROC LIFETEST data=stagec plots=survival(atrisk cb);
        time pgtime*pgstat(0);
```



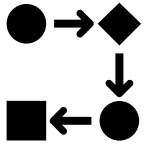

```
PROC LIFETEST data=stagec plots=survival(atrisk cb);
        time pgtime*pgstat(0);
        strata ploidy;
```



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.1



Prostate Cancer Remix: Cox regression with single variable

```
PROC PHREG data=stagec;
        class ploidy(ref='diploid');
        model pgtime*pgstat(0) = ploidy;
```

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	499.615	484.782
AIC	499.615	488.782
SBC	499.615	492.760

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
ploidy	2	13.1874	0.0014

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
ploidy	aneuploid	1	1.45706	0.46953	9.6299	0.0019	4.293	ploidy aneuploid
ploidy	tetraploid	1	1.04200	0.32634	10.1951	0.0014	2.835	ploidy tetraploid

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.2

Prostate Cancer Remix: Cox regression with multiple variables

```
PROC PHREG data=stagec;
        class eet ploidy(ref='diploid');
        model pgtime*pgstat(0)=  eet grade gleason ploidy age g2;
```

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	443.056	395.635
AIC	443.056	409.635
SBC	443.056	422.878

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
eet	1	0.0535	0.8170
grade	1	9.8729	0.0017
gleason	1	1.4413	0.2299
ploidy	2	7.1084	0.0286
age	1	0.3852	0.5348
g2	1	4.3388	0.0373

```
PROC PHREG data=stagec;
        class eet grade gleason ploidy(ref='diploid');
        model pgtime*pgstat(0)=  eet grade gleason ploidy age g2;
```

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	443.056	387.063
AIC	443.056	415.063
SBC	443.056	441.548

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
eet	1	0.2135	0.6440
grade	3	6.8078	0.0783
gleason	6	8.9328	0.1774
ploidy	2	9.2795	0.0097
age	1	0.0547	0.8150
g2	1	3.7837	0.0518

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

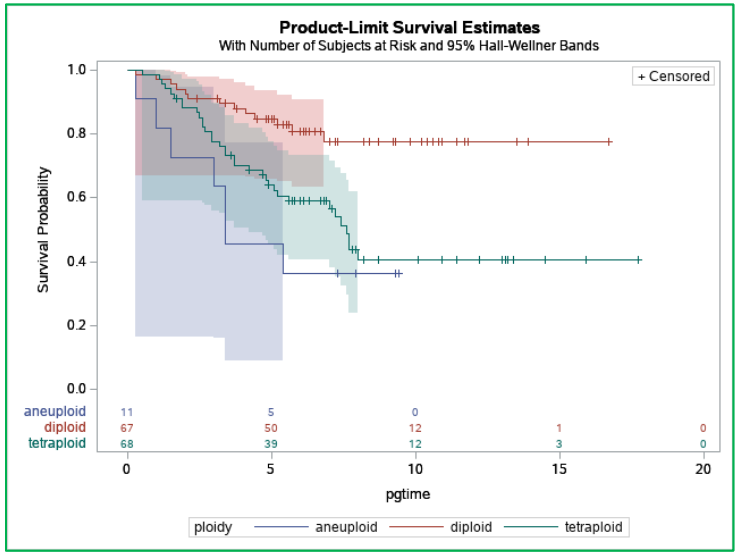
UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.2

Prostate Cancer Remix: Cox regression with multiple variables

```
PROC PHREG data=stagec;
        class grade ploidy(ref='ploidy');
        model pgtime*pgstat(0)=grade ploidy g2;
```

Compare to survival curves by ploidy category



Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	465.800	415.848
AIC	465.800	427.848
SBC	465.800	439.439

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
grade	3	33.6620	<.0001
ploidy	2	10.0724	0.0065
g2	1	4.8779	0.0272

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
grade	1	1	-15.99623	787.25505	0.0004	0.9838	0.000	grade 1
grade	2	1	-3.55911	0.61481	33.5124	<.0001	0.028	grade 2
grade	3	1	-2.21068	0.51905	18.1396	<.0001	0.110	grade 3
ploidy	aneuploid	1	1.19649	0.60057	3.9691	0.0463	3.308	ploidy aneuploid
ploidy	tetraploid	1	1.17575	0.40727	8.3342	0.0039	3.241	ploidy tetraploid
g2		1	-0.05662	0.02564	4.8779	0.0272	0.945	

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

Step-by-step Example 2.3

Prostate Cancer Remix: Parametric models [23]

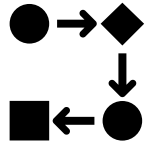
```
PROC LIFEREG data=stagec;
        class ploidy;
        model pgtime*pgstat(0)=ploidy/dist=lnormal;
PROC LIFEREG data=stagec;
        class ploidy;
        model pgtime*pgstat(0)=ploidy/dist=logistic;
PROC LIFEREG data=stagec;
        class ploidy;
        model pgtime*pgstat(0)=ploidy/dist=llogistic;
PROC LIFEREG data=stagec;
        class ploidy;
        model pgtime*pgstat(0)=ploidy/dist=gamma;
PROC LIFEREG data=stagec;
        class ploidy;
        model pgtime*pgstat(0)=ploidy/dist=exponential;
PROC LIFEREG data=stagec;
        class ploidy;
        model pgtime*pgstat(0)=ploidy/dist=weibull;
```

Distribution	AICc
lnormal	403.379
logistic	461.343
llogistic	404.043
gamma	405.514
exponential	406.302
weibull	295.114

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

Step-by-step Example 2.3



Prostate Cancer Remix: Parametric models

```
PROC LIFEREG data=stagec;
        class ploidy grade;
        model pgtime*pgstat(0)=ploidy grade g2/dist=weibull;
```

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ploidy	2	8.9327	0.0115
grade	3	33.6397	<.0001
g2	1	5.4261	0.0198

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-0.6139	0.5669	-1.7250	0.4972	1.17	0.2789
ploidy	aneuploid	1	-0.0341	0.5344	-1.0816	1.0134	0.00	0.9491
ploidy	diploid	1	0.9880	0.3642	0.2742	1.7019	7.36	0.0067
ploidy	tetraploid	0	0.0000	.	.	.	.	.
grade	1	1	21.0960	61201.47	-119932	119973.8	0.00	0.9997
grade	2	1	2.9549	0.5099	1.9555	3.9542	33.59	<.0001
grade	3	1	1.8259	0.4074	1.0275	2.6244	20.09	<.0001
grade	4	0	0.0000	.	.	.	.	.
g2		1	0.0524	0.0225	0.0083	0.0965	5.43	0.0198
Scale		1	0.8300	0.0978	0.6589	1.0456		
Weibull Shape		1	1.2048	0.1419	0.9564	1.5177		

Compared to semi-parametric Cox regression outcomes

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
grade	1	1	-15.99623	787.25505	0.0004	0.9838	0.000	grade 1
grade	2	1	-3.55911	0.61481	33.5124	<.0001	0.028	grade 2
grade	3	1	-2.21068	0.51905	18.1396	<.0001	0.110	grade 3
ploidy	aneuploid	1	1.19649	0.60057	3.9691	0.0463	3.308	ploidy aneuploid
ploidy	tetraploid	1	1.17575	0.40727	8.3342	0.0039	3.241	ploidy tetraploid
g2		1	-0.05662	0.02564	4.8779	0.0272	0.945	

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.4

Prostate Cancer: Bayesian Analysis

```
DATA stagec; set stagec;
            if ploidy='diploid' then do; psort=3; end;
            else if ploidy='tetraploid' then do; psort=2; end;
            else do; psort=1; end;
PROC SORT data=stagec;
            by psort;

PROC LIFEREG data=stagec order=data;
            class ploidy;
            model pgtime*pgstat(0)=ploidy/dist=weibull;
```

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ploidy	2	11.4246	0.0033

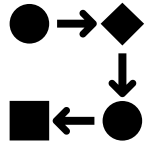
Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	3.4694	0.3176	2.8469	4.0919	119.32	<.0001
ploidy	aneuploid	1	-1.4734	0.4755	-2.4053	-0.5414	9.60	0.0019
ploidy	tetraploid	1	-0.9535	0.3346	-1.6093	-0.2977	8.12	0.0044
ploidy	diploid	0	0.0000	.	.	.	.	.
Scale		1	0.9616	0.1141	0.7621	1.2135		
Weibull Shape		1	1.0399	0.1234	0.8241	1.3122		

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.4



Prostate Cancer: Bayesian Analysis

```
PROC LIFEREG data=stagec order=data;
        class ploidy;
        model pgtime*pgstat(0)=ploidy/dist=weibull;
        bayes seed=100 outpost=post nbi=2000 nmc=10000 thin=2
                coeffprior=normal(var=1E6)
                scaleprior=gamma(shape=1E-4, iscale=1E-4);
```

Compared to parametric non-Bayesian model outcomes

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > Chi Sq
Intercept		1	3.4694	0.3176	2.8469	4.0919	119.32	<.0001
ploidy	aneuploid	1	-1.4734	0.4755	-2.4053	-0.5414	9.60	0.0019
ploidy	tetraploid	1	-0.9535	0.3346	-1.6093	-0.2977	8.12	0.0044
ploidy	diploid	0	0.0000	.	.	.	.	.
Scale		1	0.9616	0.1141	0.7621	1.2135		
Weibull Shape		1	1.0399	0.1234	0.8241	1.3122		

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits	
Intercept		1	3.4694	0.3176	2.8469	4.0919
ploidy	aneuploid	1	-1.4734	0.4755	-2.4053	-0.5414
ploidy	tetraploid	1	-0.9535	0.3346	-1.6093	-0.2977
ploidy	diploid	0	0.0000	.	.	.
Scale		1	0.9616	0.1141	0.7621	1.2135
Weibull Shape		1	1.0399	0.1234	0.8241	1.3122

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
Intercept	5000	3.5838	0.3476	3.3432	3.5540	3.7939
ploidyaneuploid	5000	-1.5023	0.5117	-1.8350	-1.4973	-1.1693
ploidytetraploid	5000	-1.0183	0.3626	-1.2521	-0.9964	-0.7586
Scale	5000	1.0166	0.1271	0.9245	1.0063	1.0924

Posterior Intervals

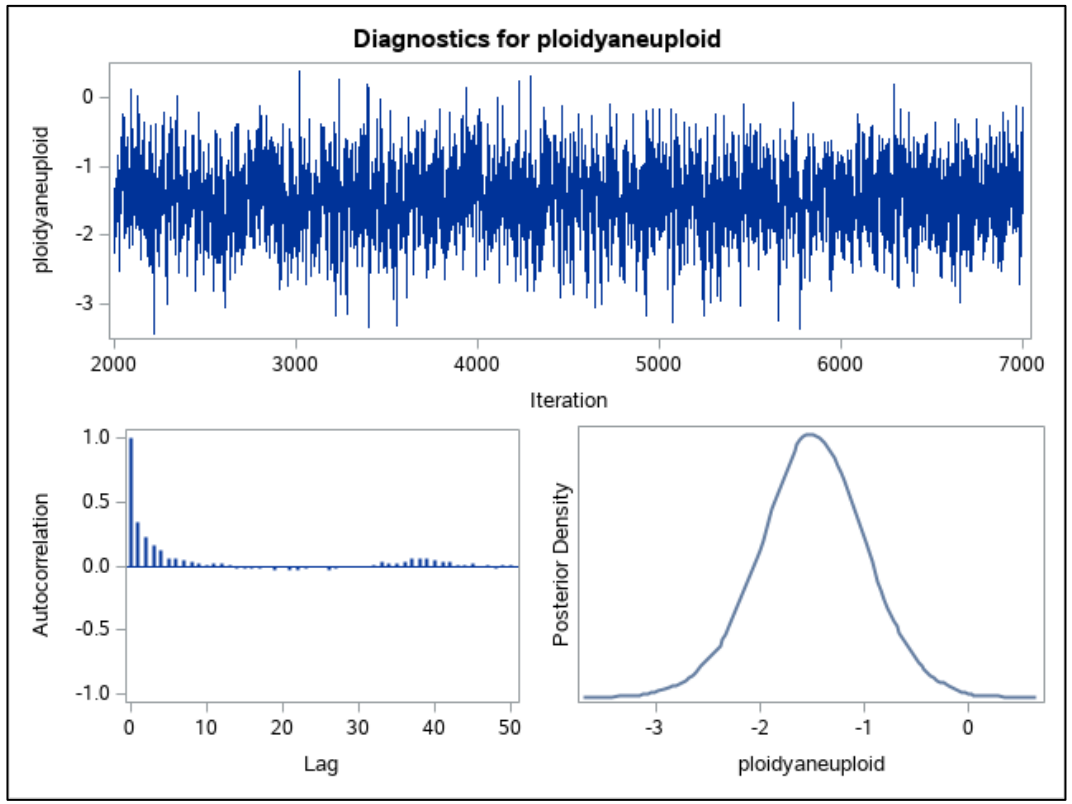
Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Intercept	0.050	2.9829	4.3633	2.9157	4.2751
ploidyaneuploid	0.050	-2.5412	-0.5012	-2.5623	-0.5273
ploidytetraploid	0.050	-1.7861	-0.3781	-1.7520	-0.3575
Scale	0.050	0.8019	1.2989	0.7823	1.2677

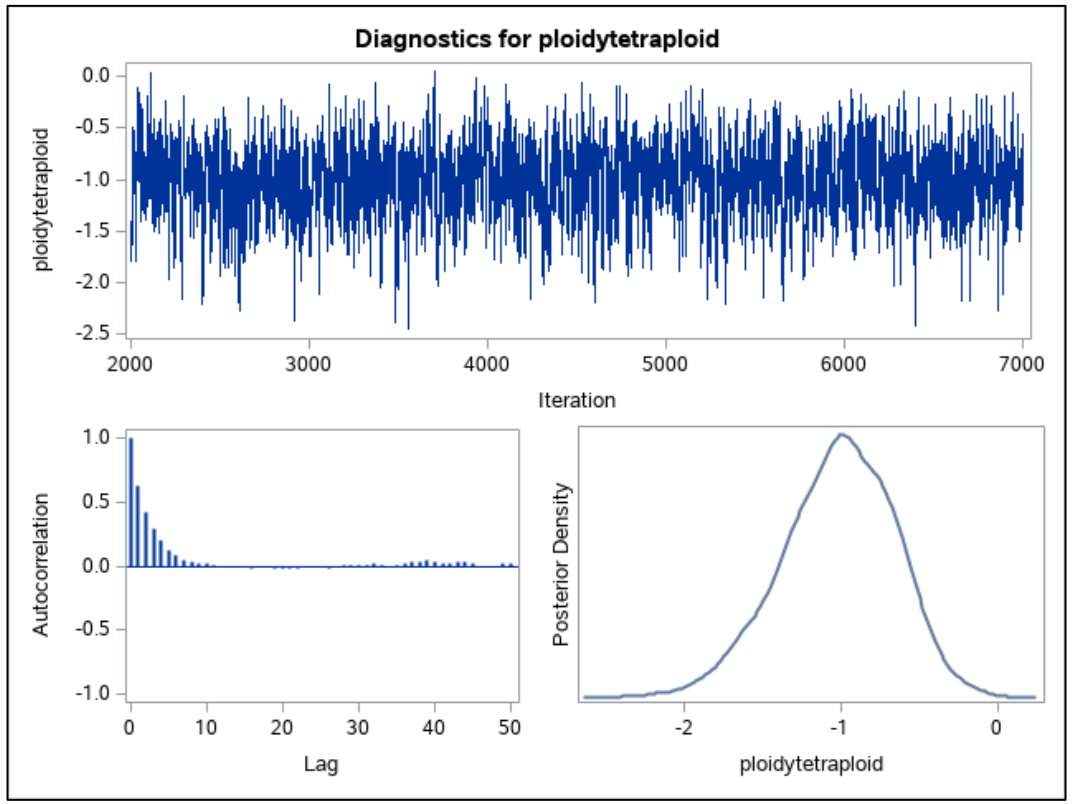
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

Step-by-step Example 2.4

Prostate Cancer: Bayesian Analysis

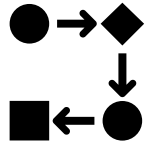




DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.5



Heart Transplant: Time-Dependent Variables [24]

```
PROC PHREG data=Heart;
          model Time*Status(0)=XStatus Acc_Age;
          if (WaitTime=. or Time < WaitTime) then XStatus=0.;
          else XStatus=1.0;

PROC PHREG data= Heart;
  model Time*Status(0)= XStatus XAge XScore;
  where NotTyped ^= 'y';
  if (WaitTime = . or Time < WaitTime) then do;
    XStatus=0.;
    XAge=0.;
    XScore= 0.;
  end;
  else do;
    XStatus= 1.0;
    XAge= Xpl_Age;
    XScore= Mismatch;
  end;
```

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	596.651	591.292
AIC	596.651	595.292
SBC	596.651	599.927

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	561.680	551.874
AIC	561.680	557.874
SBC	561.680	564.662

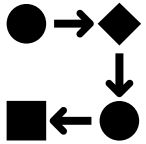
Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
XStatus	1	-0.06720	0.30594	0.0482	0.8261	0.935
Acc_Age	1	0.03158	0.01446	4.7711	0.0289	1.032

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
XStatus	1	-3.19837	1.18746	7.2547	0.0071	0.041
XAge	1	0.05544	0.02263	6.0019	0.0143	1.057
XScore	1	0.44490	0.28001	2.5245	0.1121	1.560

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.6



Leukemia: Competing Risks [25]

```
PROC PHREG data=bmt; *for hazard of relapse;
        class Group / order=internal ref=first param=glm;
        model T*Status(0,2) = Group logWaitTime;
        hazardratio 'Cause-Specific Hazards' Group / diff=pairwise;
  *Cause-specific Hazard (competing events are treated as censored events)
  *Status: 0=Censored, 1=Relapse, 2=Die (w/o relapse)
```

```
DATA risk;
        logWaitTime=5.2;
        Group=1; output;
        Group=2; output;
        Group=3; output;
        format Group DiseaseGroup.;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
Group	AML-Low Risk	1	-1.07299	0.46245	5.3836	0.0203	0.342	Group AML-Low Risk
Group	AML-High Risk	1	0.55111	0.36464	2.2842	0.1307	1.735	Group AML-High Risk
Group	ALL	0	0	.	.	.	.	Group ALL
logWaittime		1	-0.23061	0.19440	1.4071	0.2355	0.794	

Cause-Specific Hazards: Hazard Ratios for Group

Description	Point Estimate	95% Wald Confidence Limits	
Group AML-Low Risk vs AML-High Risk	0.197	0.088	0.441
Group AML-High Risk vs AML-Low Risk	5.074	2.268	11.353
Group AML-Low Risk vs ALL	0.342	0.138	0.847
Group ALL vs AML-Low Risk	2.924	1.181	7.238
Group AML-High Risk vs ALL	1.735	0.849	3.546
Group ALL vs AML-High Risk	0.576	0.282	1.178

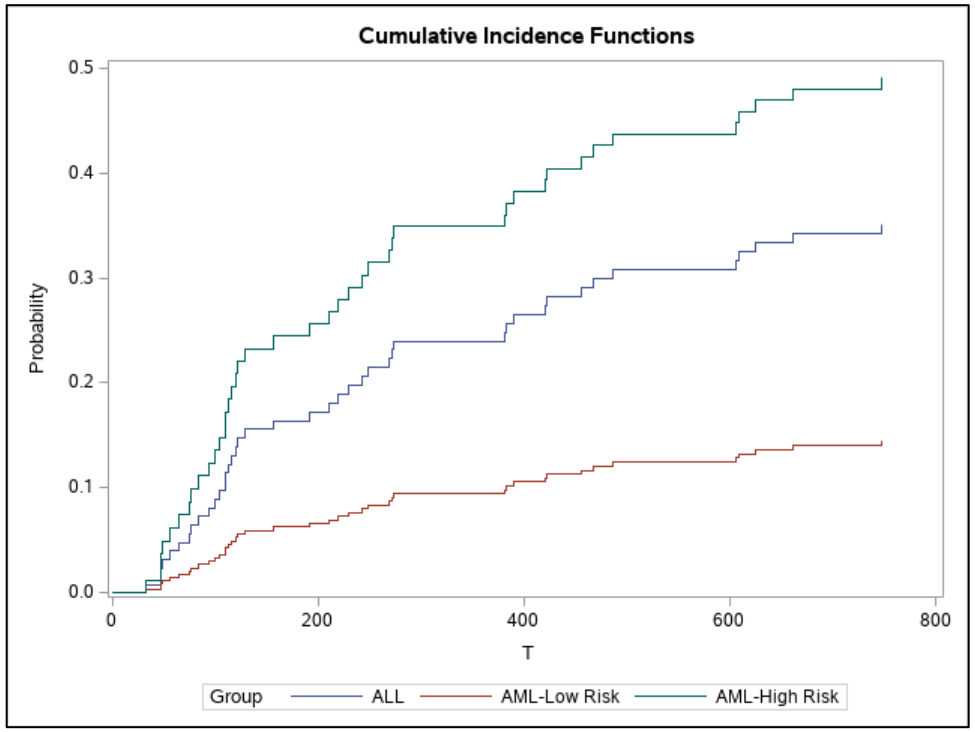
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2.6

Leukemia: Competing Risks

```
PROC PHREG data=bmt plots(overlay=stratum)=cif;
        class Group / order=internal ref=first param=glm;
        model T*Status(0) = Group logWaitTime / eventcode=1;
                hazardratio 'Subdistribution Hazards' Group /diff=pairwise;
        baseline covariates=risk out=_null_ /rowid=Group;
```



Subdistribution Hazards: Hazard Ratios for Group			
Description	Point Estimate	95% Wald Confidence Limits	
Group AML-Low Risk vs AML-High Risk	0.231	0.106	0.503
Group AML-High Risk vs AML-Low Risk	4.323	1.990	9.394
Group AML-Low Risk vs ALL	0.362	0.155	0.843
Group ALL vs AML-Low Risk	2.765	1.186	6.445
Group AML-High Risk vs ALL	1.564	0.763	3.203
Group ALL vs AML-High Risk	0.640	0.312	1.310

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
Group	AML-Low Risk	1	-1.01701	0.43177	5.5481	0.0185	0.362	Group AML-Low Risk
Group	AML-High Risk	1	0.44702	0.36591	1.4925	0.2218	1.564	Group AML-High Risk
Group	ALL	0	0	.	.	.	.	Group ALL
logWaittime		1	-0.28540	0.19563	2.1283	0.1446	0.752	

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Assessment 2

qualtrics XM

UNIVERSITY OF NORTH DAKOTA UND

https://und.qualtrics.com/jfe/form/SV_aY2HKfZSPSx8oJ0

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Caveats and Concerns 

- We just scratched the surface on many of the examples
  - Lots of tweaks to exploration, model parameters, and graphing
  - Didn’t cover landmark analysis, conditional survival, or logistic-style regression
- Model fitting and interpretation are important to get right
  - Different procedures will have different outputs
  - Same work that goes into multiple regression or generalized linear models applies
  - Even parametric models with a flexible distribution may be a poor fit
  - Splines or fractional polynomials may be needed [4]
- As always, start simple and build from there

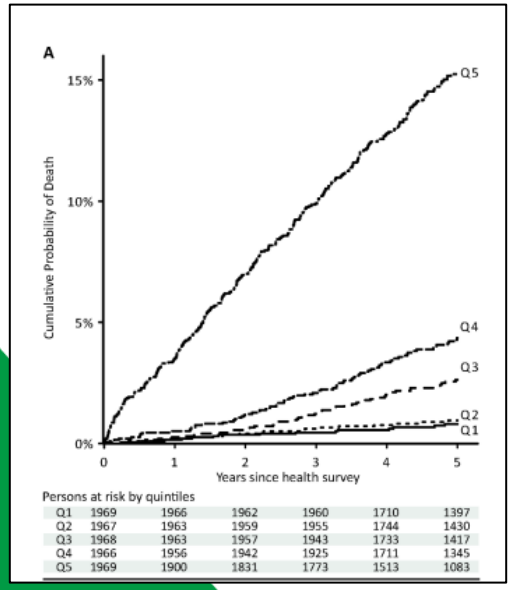
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

Real World Examples

Fischer K, Kettunen J, Würtz P, Haller T, Havulinna AS, et al. (2014) Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality: An Observational Study of 17,345 Persons. PLOS Medicine 11(2): e1001606. https://doi.org/10.1371/journal.pmed.1001606 **[26]**

Early identification of ambulatory persons at high short-term risk of death could benefit targeted prevention. To identify biomarkers for all-cause mortality and enhance risk prediction, we conducted high-throughput profiling of blood specimens in two large population-based cohorts



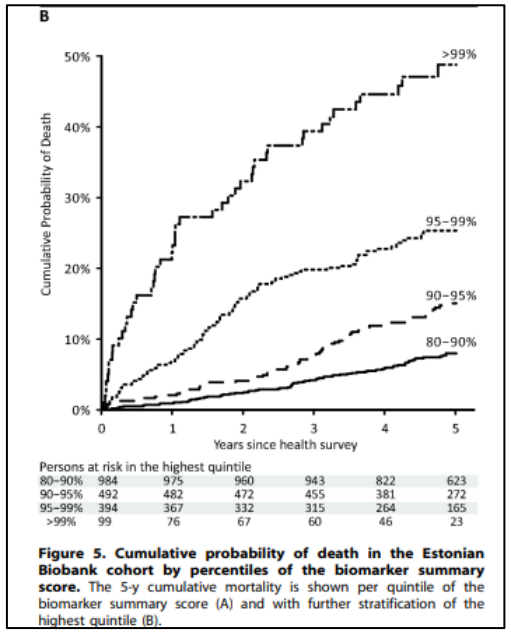


Figure 5. Cumulative probability of death in the Estonian Biobank cohort by percentiles of the biomarker summary score. The 5-y cumulative mortality is shown per quintile of the biomarker summary score (A) and with further stratification of the highest quintile (B).

Table 2. Hazard ratios for all-cause mortality derived in the Estonian Biobank cohort in the age range 25–74 y.

Variable	Prediction Model without Biomarkers			Prediction Model with Biomarkers		
	HR	95% CI	*p*-Value	HR	95% CI	*p*-Value
Female gender	0.67	0.50–0.90	0.009	0.60	0.44–0.81	0.0008
Body mass index[a]	1.05	0.91–1.21	0.52	1.05	0.92–1.20	0.48
Systolic blood pressure[a]	0.96	0.85–1.09	0.51	1.04	0.92–1.18	0.55
Fasting duration (hours)	0.99	0.96–1.02	0.47	1.00	0.97–1.03	0.96
Total cholesterol[a]	1.05	0.91–1.21	0.50	1.15	0.97–1.36	0.11
HDL-cholesterol[a]	0.81	0.69–0.95	0.01	1.07	0.92–1.24	0.37
Triglycerides[a]	0.82	0.70–0.96	0.01	0.93	0.71–1.21	0.60
Creatinine[a]	1.10	1.03–1.18	0.005	1.04	0.96–1.12	0.31
Current smoking	1.86	1.26–2.75	0.002	1.56	1.05–2.33	0.03
Smoking duration (years)[a]	1.21	1.04–1.41	0.01	1.25	1.07–1.46	0.005
Cigarettes per day[a]	0.93	0.80–1.07	0.29	0.89	0.77–1.03	0.11
Alcohol[a]	1.09	0.98–1.21	0.11	1.04	0.94–1.16	0.43
Prevalent diabetes	1.58	1.15–2.15	0.004	1.49	1.09–2.03	0.01
Prevalent cardiovascular disease	1.38	1.05–1.82	0.02	1.42	1.08–1.87	0.01
Prevalent cancer	2.15	1.51–3.05	2×10^{-5}	2.26	1.59–3.20	5×10^{-6}
Alpha-1-acid glycoprotein[a]	—	—	—	1.76	1.57–1.97	9×10^{-23}
Albumin[a]	—	—	—	0.66	0.59–0.73	4×10^{-15}
VLDL particle size[a]	—	—	—	0.74	0.58–0.94	0.01
Citrate[a]	—	—	—	1.47	1.29–1.67	5×10^{-9}

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

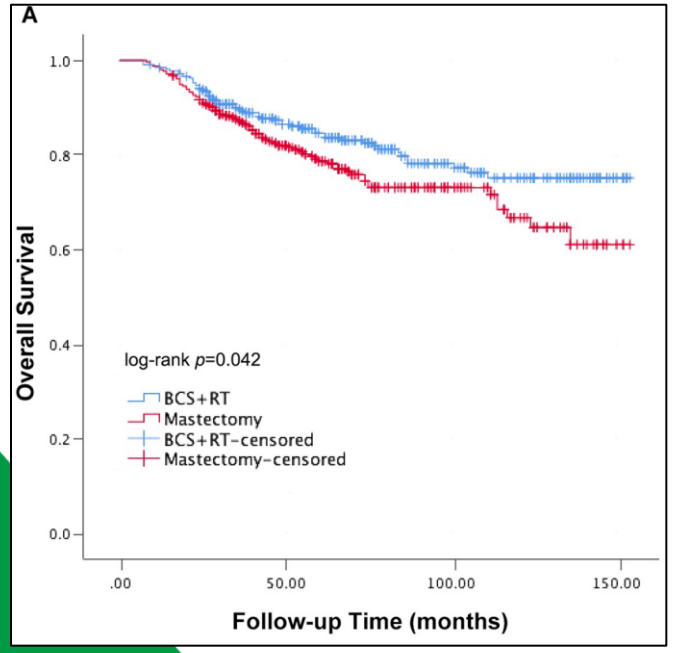
UND UNIVERSITY OF NORTH DAKOTA.

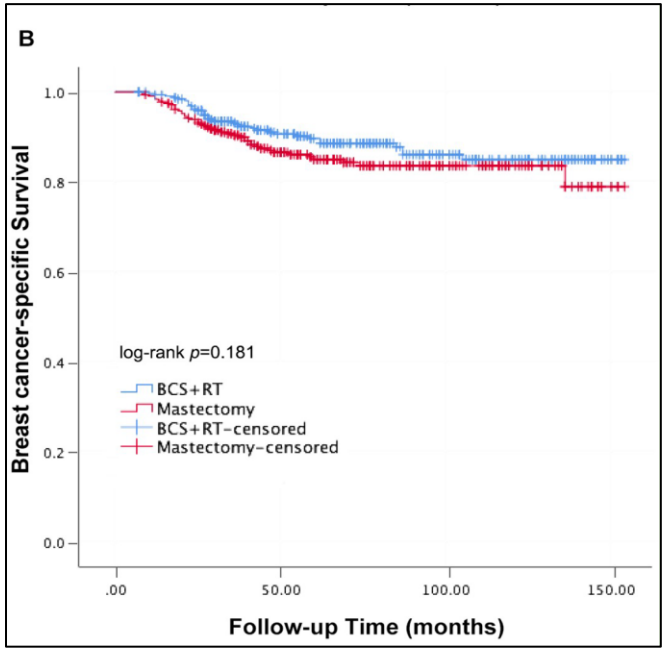
Real World Examples

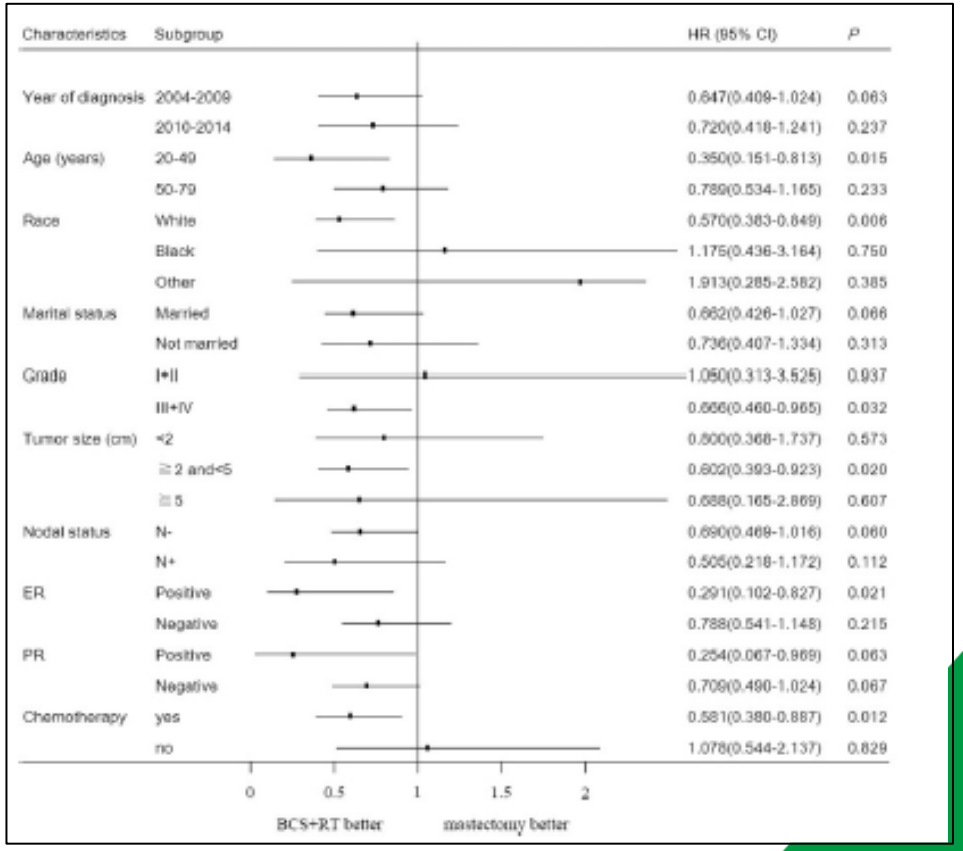
Xia LY, Xu WY, Hu QL (2021) The different outcomes between breast-conserving surgery plus radiotherapy and mastectomy in metaplastic breast cancer: A population-based study. PLOS ONE 16(9): e0256893. https://doi.org/10.1371/journal.pone.0256893 **[27]**

Metaplastic breast cancer (MBC) are rare. The survival outcomes of MBC patients after breast conserving surgery plus radiotherapy (BCS+RT) or mastectomy have not been established. The study aimed to compare survival outcomes of MBC patients subjected to BCS+RT or mastectomy therapeutic options.

The conditional landmark analysis was used to address a lead time bias among the propensity matched cohort. With the landmark, analysis was restricted to the patients who survived to 6 months without death or loss to follow-up.







Real World Examples

Gaspar MJ, Velasco T, Feito I, Alía R, Majada J (2013) Genetic Variation of Drought Tolerance in Pinus pinaster at Three Hierarchical Levels: A Comparison of Induced Osmotic Stress and Field Testing. PLOS ONE 8(11): e79094. https://doi.org/10.1371/journal.pone.0079094 **[28]**

We performed a Polyethylene glycolosmotic induced stress experiment and evaluated two common garden experiments (xeric and mesic sites) to test for survival and growth of a wide range clonal collection of Maritime pine.

To estimate the proportion of plants surviving at a given time, and hence the survival probability at that time for each clone, the Kaplan-Meier method was used as a product-limit estimator . This principle makes it possible to work with conditional and cumulative probabilities.

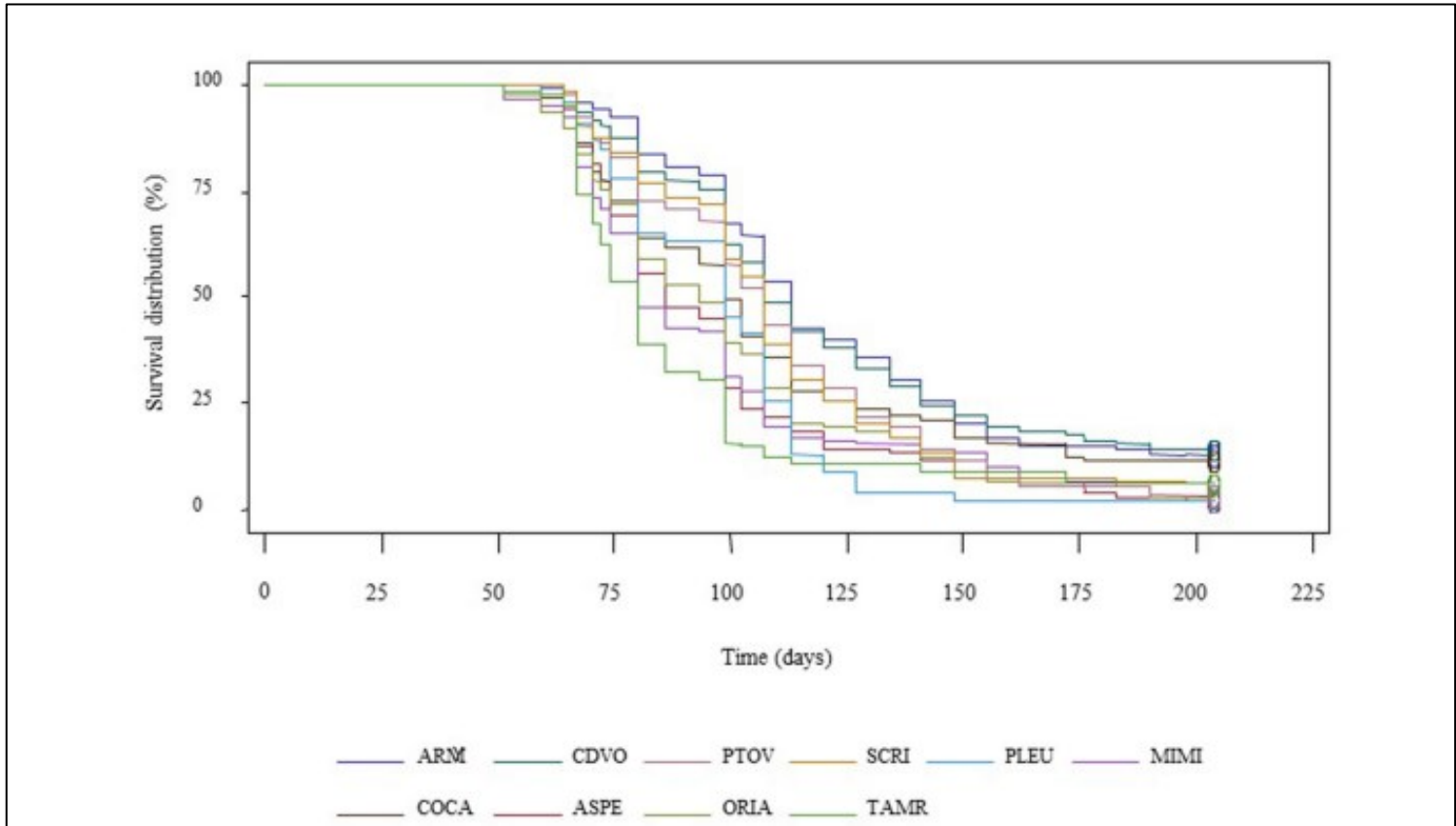
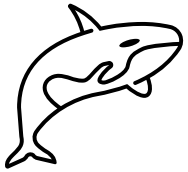


Figure 1. Distribution function of survival by provenance during the PEG-induced osmotic stress experiment, estimated by Kaplan-Meier method.

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Summary and Conclusion



- Cox regression is the survival analysis equivalent to linear regression
- Parametric models are the survival analysis equivalent to generalized linear models
- Tree based methods are an alternative to Cox regression
- There are a variety of particular methods—such as frailty models, competing risks, and time-dependent variables—that are used under specific circumstances
- At the end of the day, we can determine the effects of factors on survival times

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

References

[1] https://www.ibm.com/docs/en/spss-statistics/24.0.0?topic=option-cox-regression-analysis

[2] https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival6.html

[3] https://www.datacamp.com/community/tutorials/survival-analysis-R

[4] https://www.r-bloggers.com/2019/06/parametric-survival-modeling/

[5] http://www.stat.columbia.edu/~madigan/W2025/notes/survival.pdf

[6] https://www.xlstat.com/en/solutions/features/parametric-survival-regression-weibull-model

[7] https://data.princeton.edu/pop509/parametricsurvival.pdf

[8] https://www.math.ucsd.edu/~rxu/math284/slect4.pdf

[9] https://www.andreaperlato.com/tspost/parametric-regression-model-in-survival-analysis/

[10] https://eclass.uoa.gr/modules/document/file.php/MATH297/SAS%20labs/saslab07.pdf

[11] https://en.wikipedia.org/wiki/Survival_analysis

[12] https://riptutorial.com/r/example/13086/random-forest-survival-analysis-with-randomforestsrc

[13] https://pubmed.ncbi.nlm.nih.gov/32466712/

[14] https://www.theanalysisfactor.com/the-six-types-of-survival-analysis-and-challenges-in-learning-them/

[15] https://www.stata.com/support/faqs/statistics/multiple-failure-time-data/

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

References 2

[16] https://www.demogr.mpg.de/papers/working/wp-2003-032.pdf

[17] https://www.publichealth.columbia.edu/research/population-health-methods/competing-risk-analysis

[18] https://www.rensvandeschoot.com/tutorials/discrete-time-survival/

[19] https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html

[20] https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf

[21] https://onlinelibrary.wiley.com/doi/epdf/10.1002/cnr2.1210

[22] https://www.rdocumentation.org/packages/survival/versions/3.2-13/topics/frailty

[23] https://support.sas.com/resources/papers/proceedings10/252-2010.pdf

[24] https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_phreg_sect049.htm

[25] https://support.sas.com/rnd/app/stat/papers/2014/competingrisk2014.pdf

[26] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001606

[27] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0256893

[28] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079094

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Acknowledgements

UND
UNIVERSITY OF NORTH DAKOTA.

- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY