STATISTICAL LEARNING IN EPIDEMIOLOGY FINAL PROJECT

A study of prognostic indicators in cancer patients

Bortolussi Lorenzo, Cartago Marco, Tonet Lorenzo

THE DATASET

The data used in this project comes from the study:

"Prospective Evaluation of Prognostic Variables From Patient-Completed Questionnaires"

In this study the data was collected via a questionnaire given to patients that simultaneously entered the North Central Cancer Treatment Group, before reciving their fist dose of chemotherapy. The study began during **July 15-17 1987** and was concluded after three years.

THE DATASET

The physicians collected information in three key areas:

- **Nutritional factors**
- **Physician assesment**
- **Patient assessment**

For the 1077 patients observed, researches collected 60 variables, 36 of which were considered significant.

The dataset available to us is a subset of the original dataset, containing 228 patients and 9 variables, which can still be grouped in the previous three categories.

General patient informations:

- **inst**: The code of the institution in which the patient was hospitalized
- **time**: The survival time in days
- **status**: The censoring status
- **age**: The age in years
- **sex**: Patient sex

Food intake:

- **meal.cal**: Calories consumed at the meal before hospitalization
- **wt.loss**: Weight loss in pounds in the last six months

Health metrics:

- **ph.ecog**: ECOG performance score as rated by the physician, divided in: asymptomatic; symptomatic but ambulatory; in bed <50% of the day; in bed > 50% of the day but not bedbound; bedbound
- **ph.karno**: Karnofsky performance score (bad=0-good=100) rated by physician
- **pat.karno**: Karnofsky performance score as rated by the patient itself

OUR QUESTION

Our question concerns the creation of a prognostic model and the identification of possible prognostic factors.

One thing that interested us in particular was to what extent is patient self-evaluation a good prognostic factor (**pat.karno**), compared to physician evaluation (**ph.karno**)

PROJECT STRUCTURE

Initial data analysis and exploration

- NA distribution
- Univariate and bivariate distributions

Survival analysis

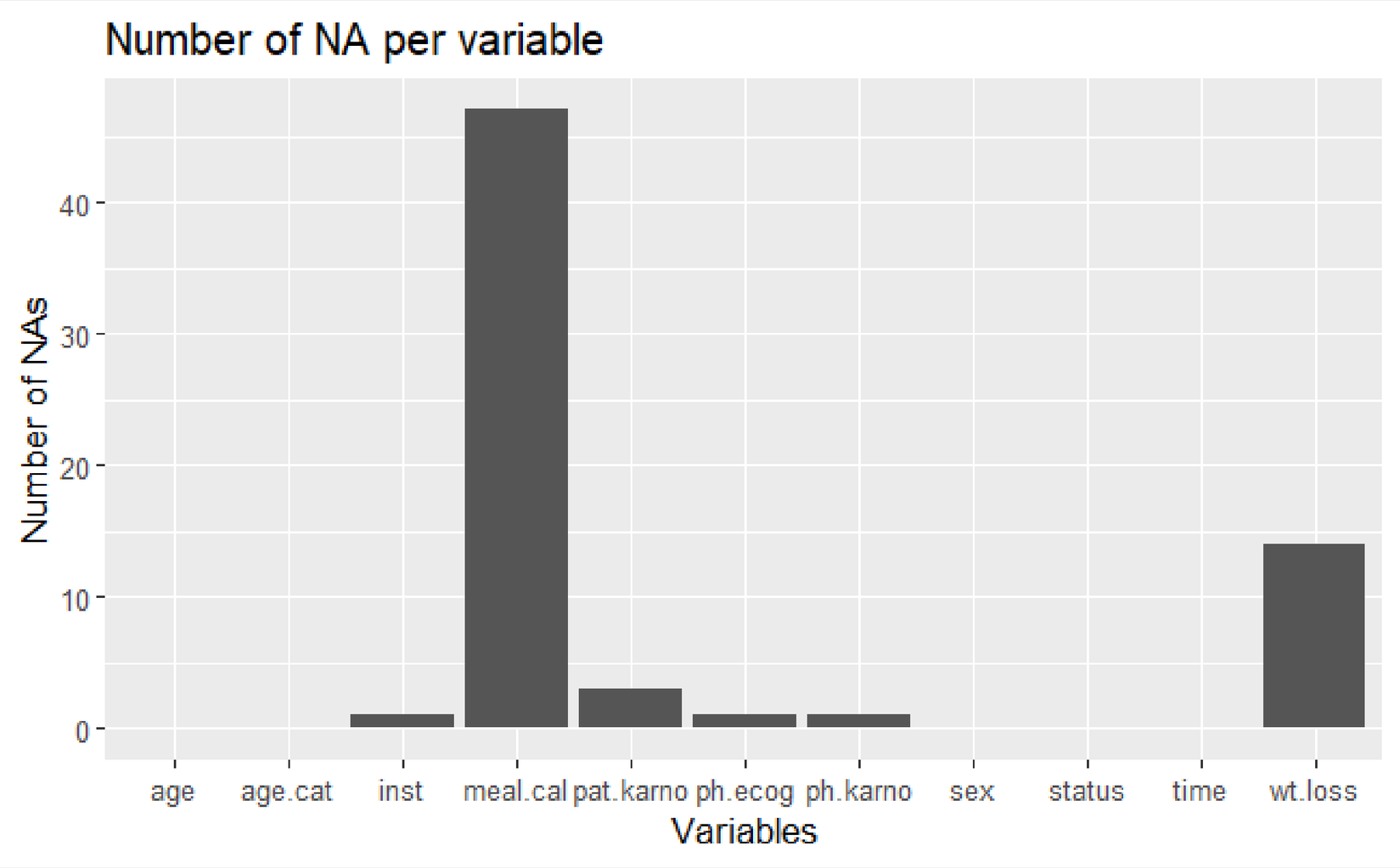
- Using Kaplan-Meier estimators and Cox models to compare different prognostic factors

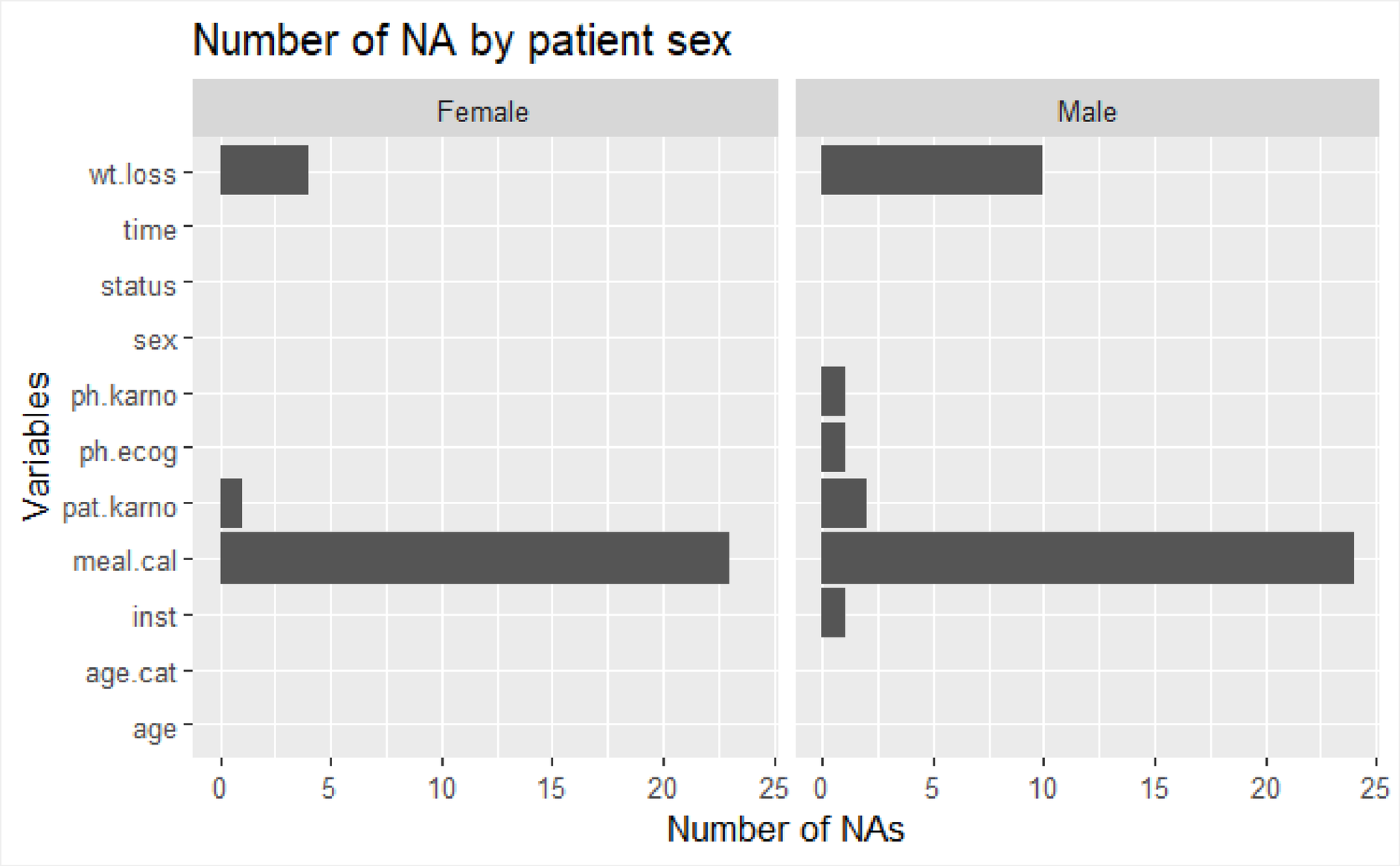
Model assessment

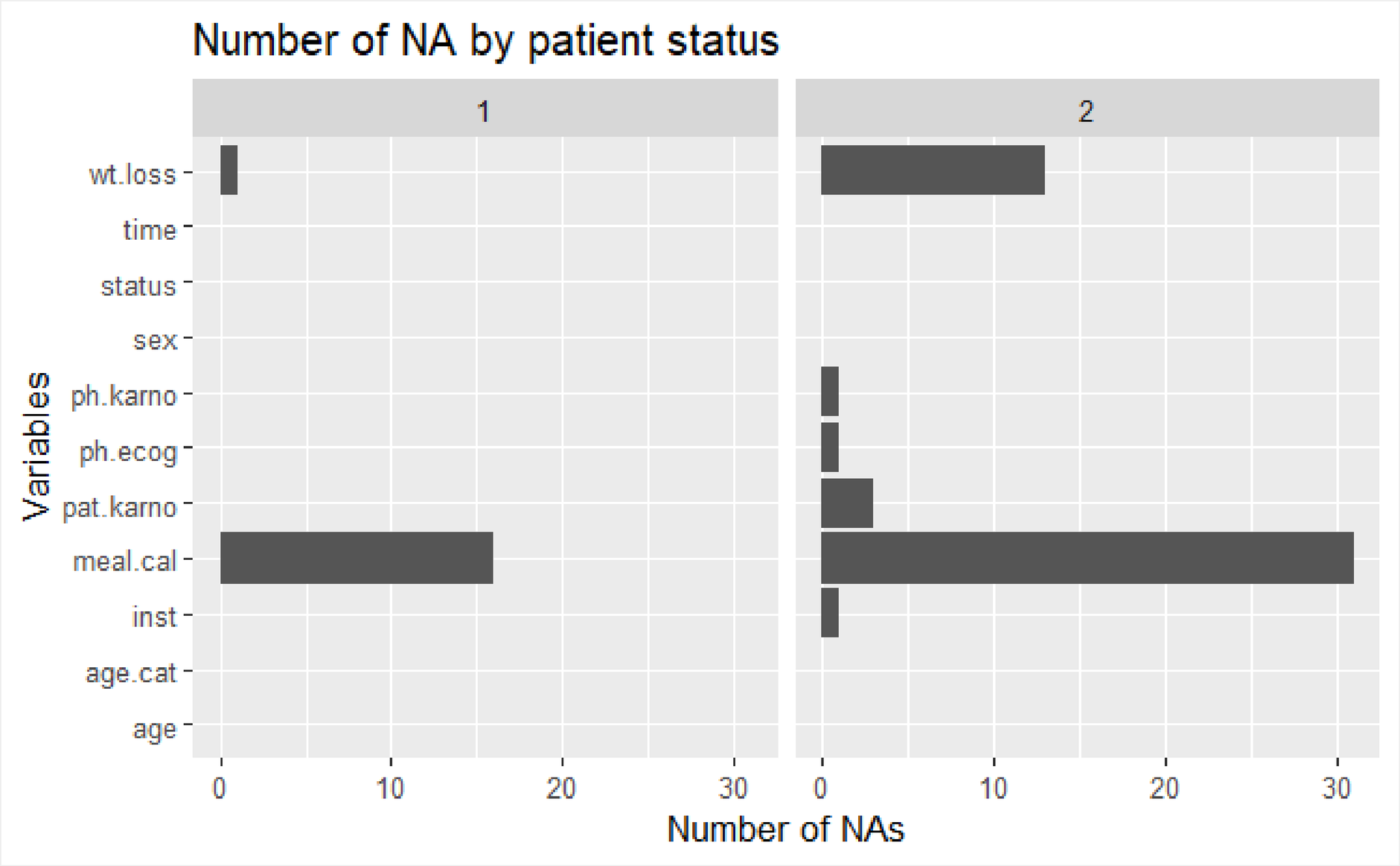
- Discrimination and calibration of the models

INITIAL DATA ANALYISIS

HOW ARE NAS DISTRIBUTED?




Number of NA by patient sex
Female
Male
wt.loss
time
status
sex
ph.karno
ph.ecog
pat.karno
meal.cal
inst
age.cat
age
Variables
0
5
10
15
20
25
Number of NAs


Number of NA by patient status
1
2
wt.loss
time
status
sex
ph.karno
ph.ecog
pat.karno
meal.cal
inst
age.cat
age
Variables
0
10
20
30
Number of NAs

UNIVARIABLE DISTRIBUTIONS

STATUS:

Censored : 26.6%

Dead : 73.4%

AGE CATEGORY:

<50 y : 11.4%

50 - 70 y : 68.4%

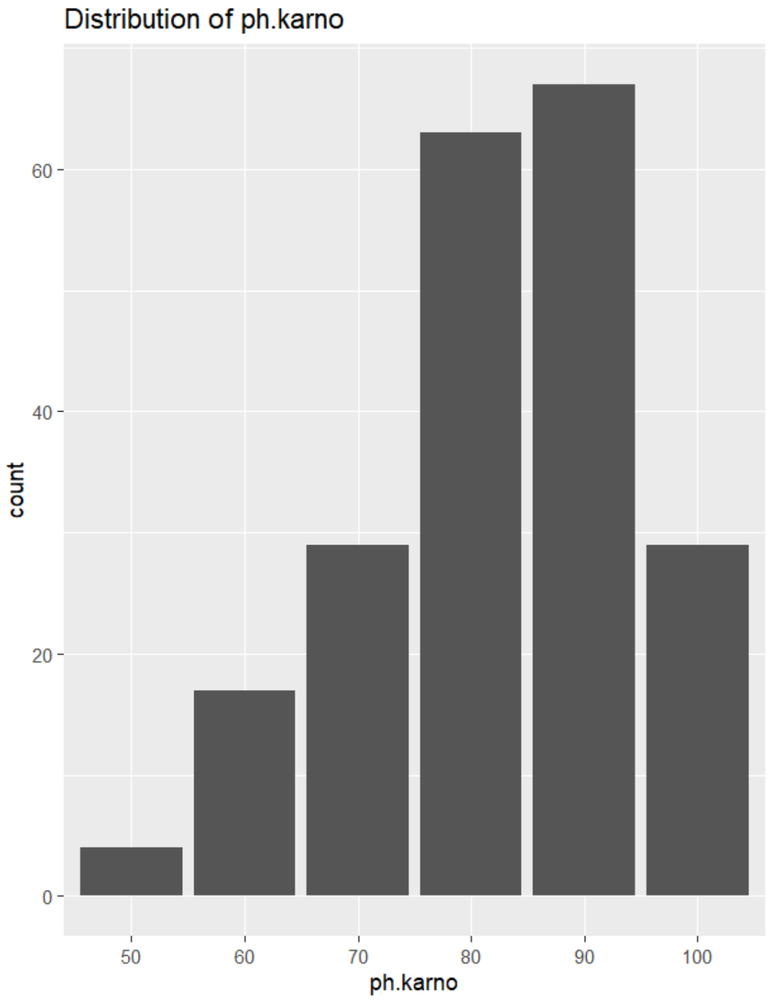
>70 : 20.2%

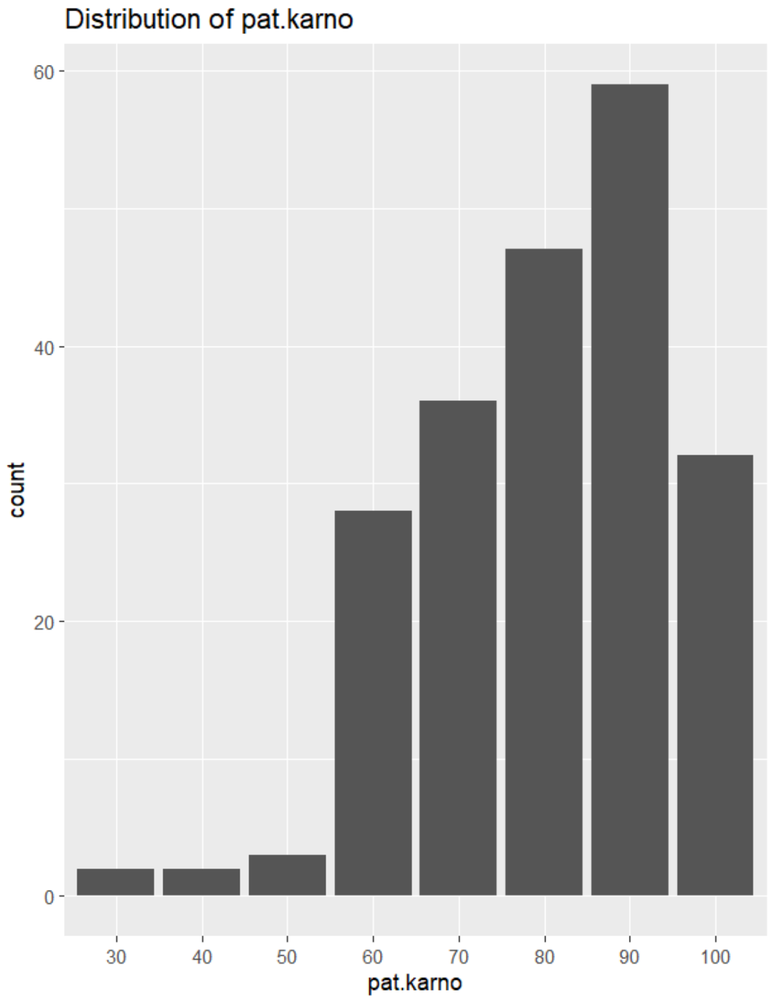
SEX:

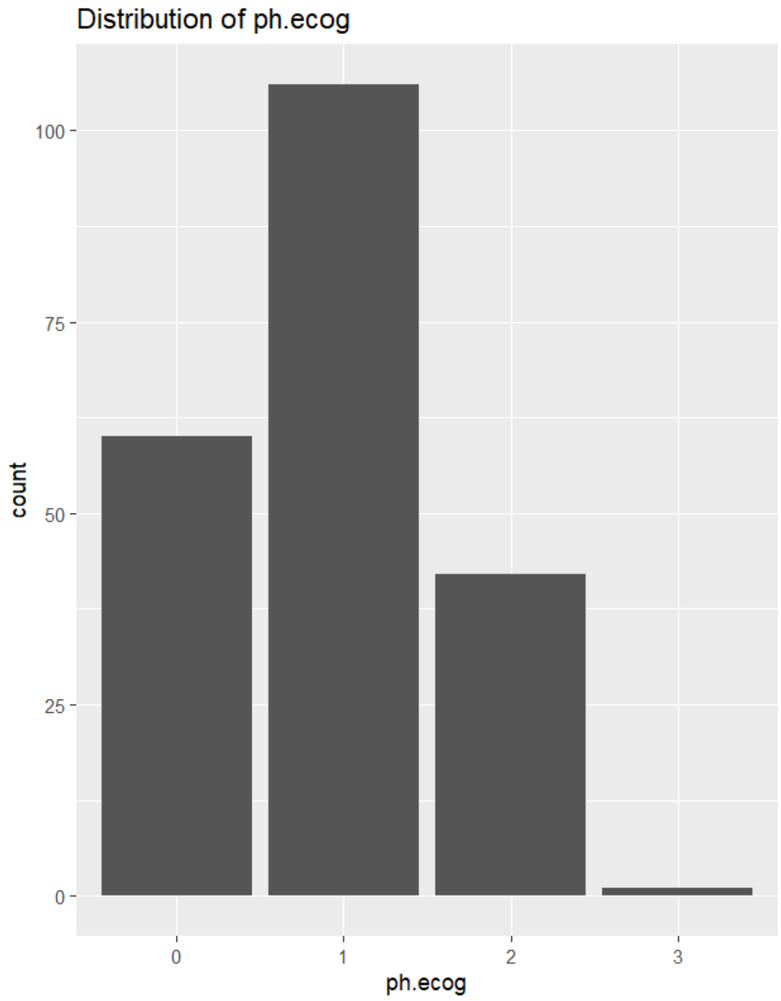
Male : 60.5%

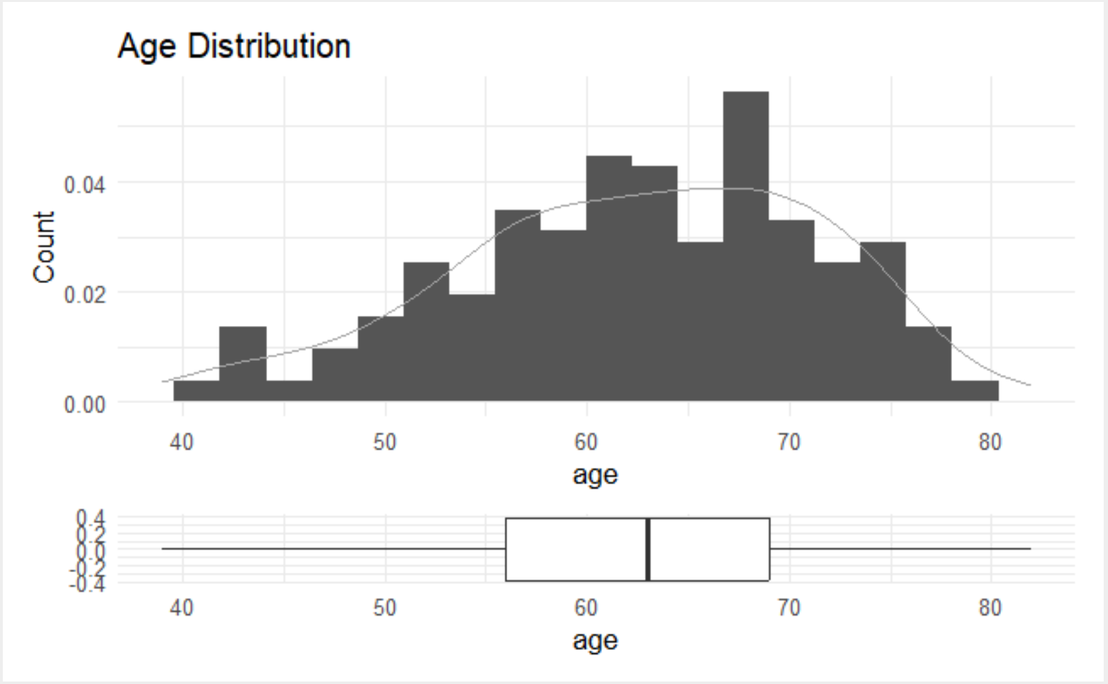
Female : 39.5%

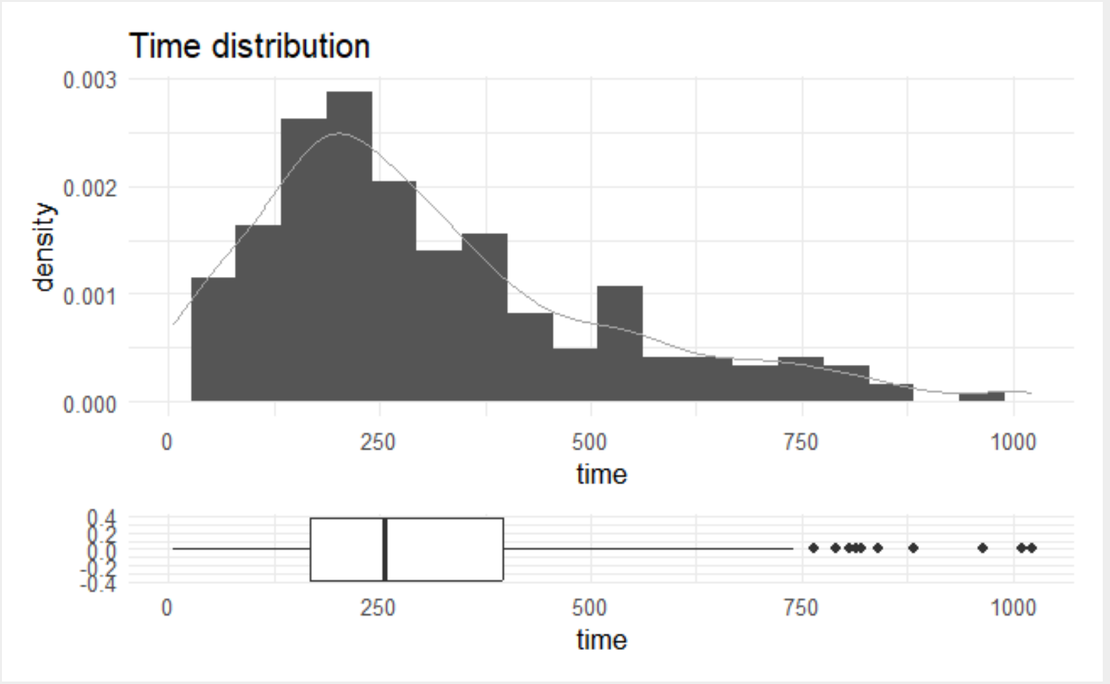
UNIVARIABLE DISTRIBUTIONS

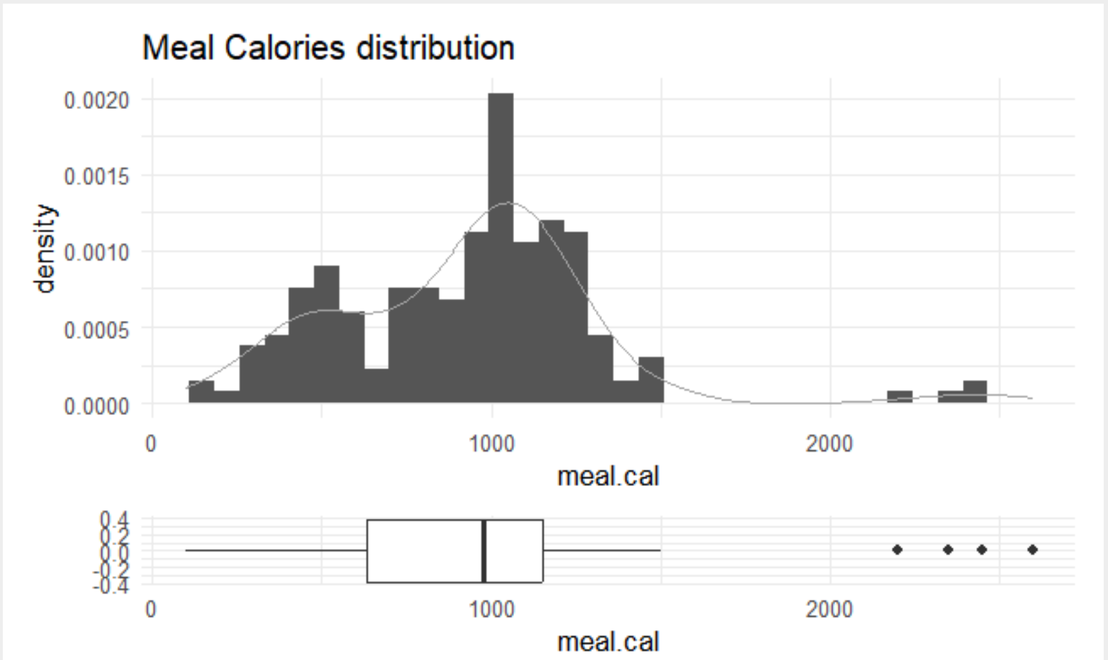


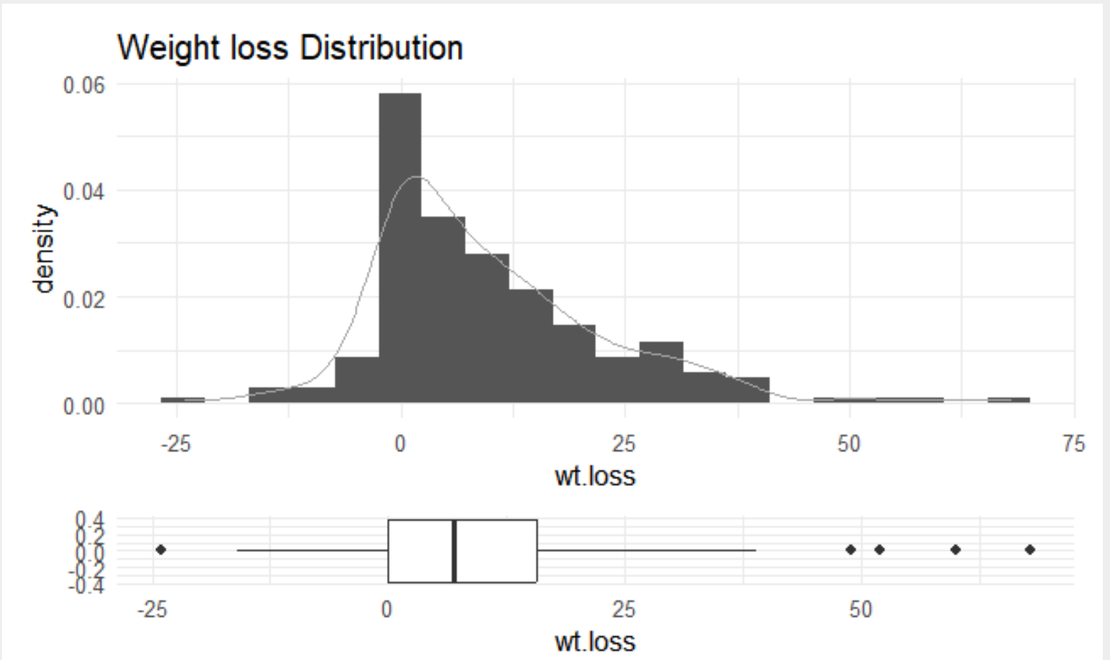




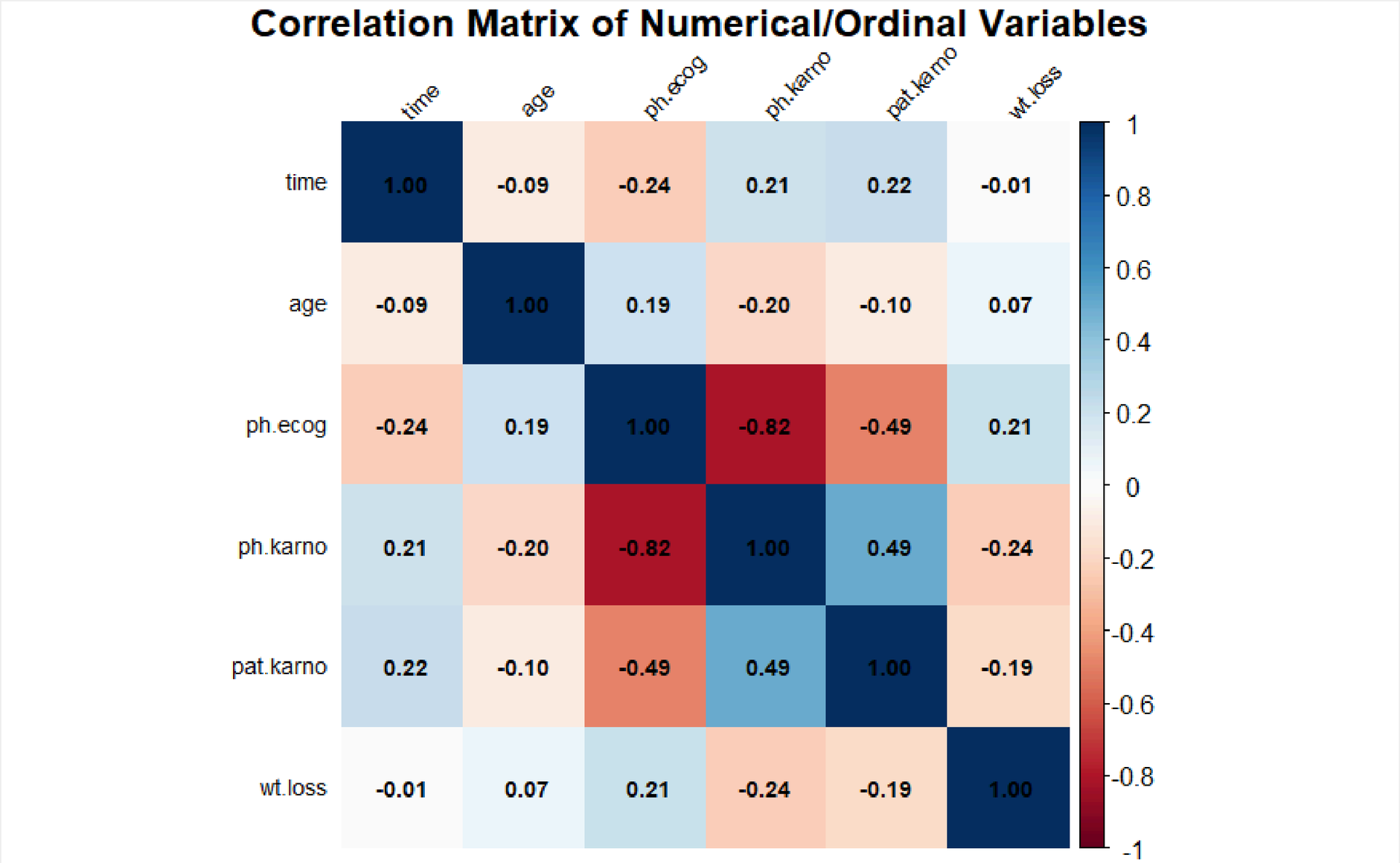

Age Distribution
Count
0.04
0.02
0.00
40
50
60
70
80
age
0.4
0.2
0.0
-0.2
-0.4
40
50
60
70
80
age

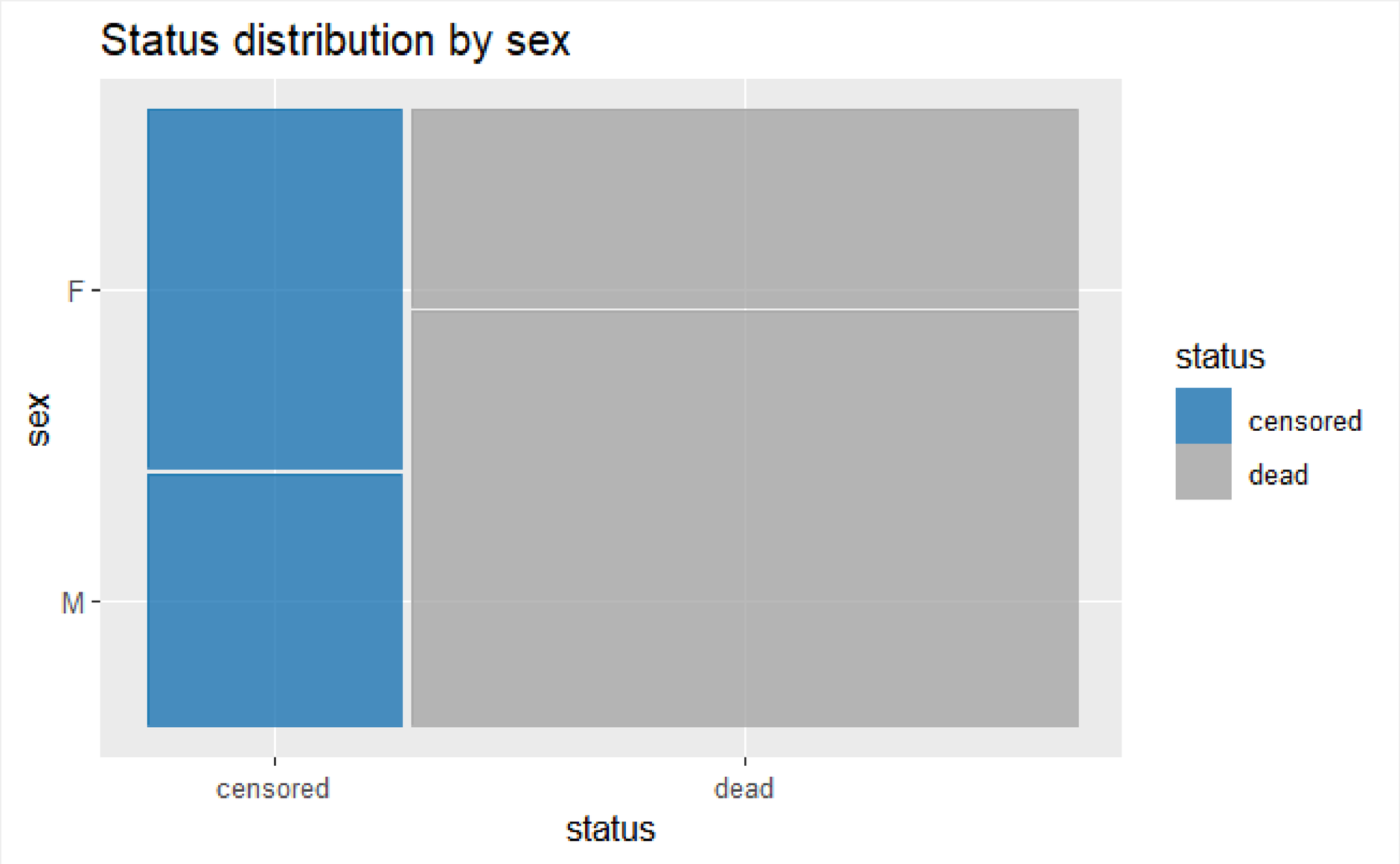

Time distribution
density
0.003
0.002
0.001
0.000
0
250
500
750
1000
time
0.4
0.2
0.0
-0.2
-0.4
0
250
500
750
1000
time

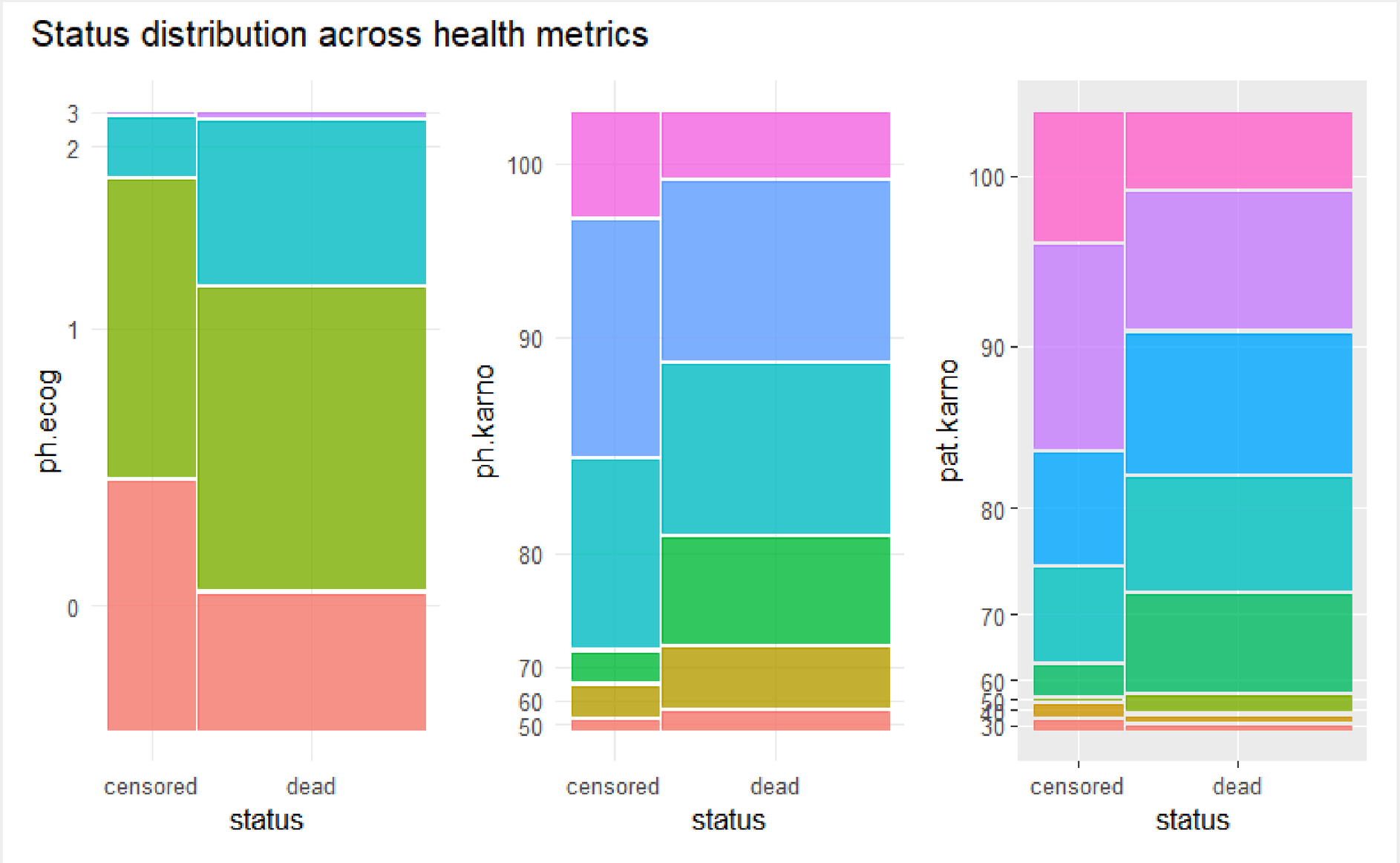

Meal Calories distribution
density
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
meal.cal
0.4
0.2
0.0
-0.2
-0.4
0
1000
2000
meal.cal

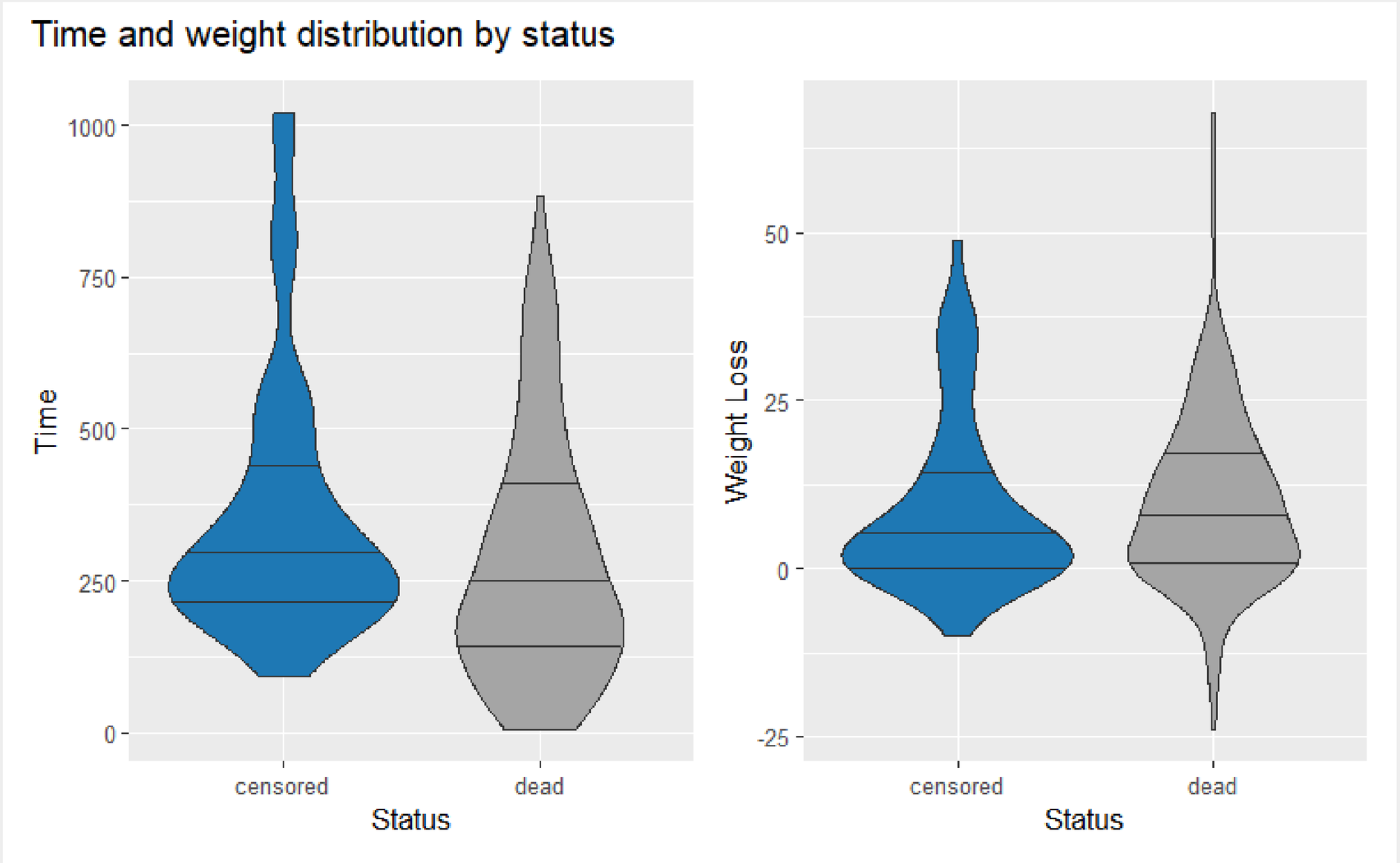

Weight loss Distribution
density
0.06
0.04
0.02
0.00
-25
0
25
50
75
wt.loss
0.4
0.2
0.0
-0.2
-0.4
-25
0
25
50
wt.loss

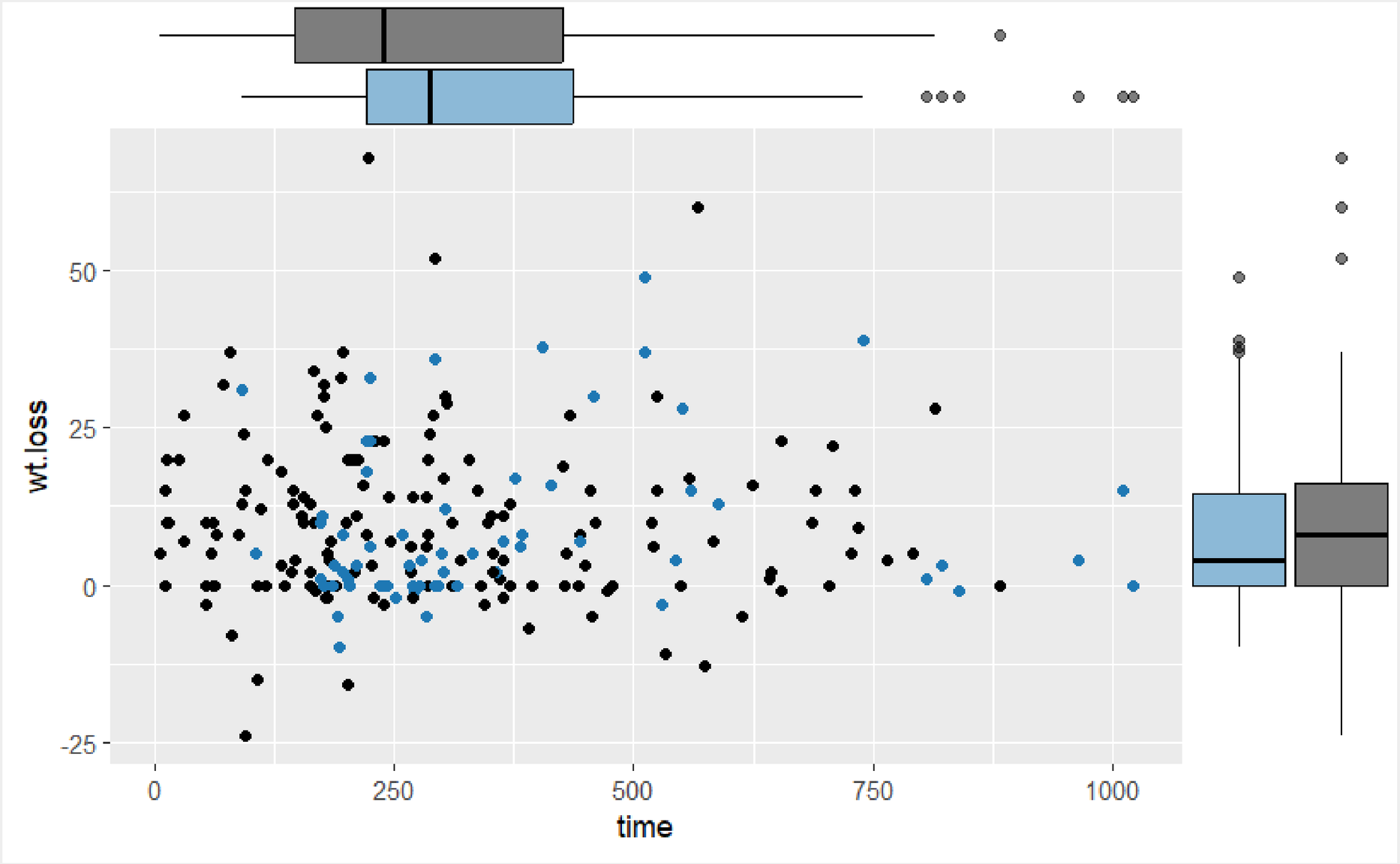
BIVARIATE ANALYSIS


Correlation Matrix of Numerical/Ordinal Variables
time
age
ph.ecog
ph.karno
pat.karno
wt.loss
1.00 -0.09 -0.24 0.21 0.22 -0.01
-0.09 1.00 0.19 -0.20 -0.10 0.07
-0.24 0.19 1.00 -0.82 -0.49 0.21
0.21 -0.20 -0.82 1.00 0.49 -0.24
0.22 -0.10 -0.49 0.49 1.00 -0.19
-0.01 0.07 0.21 -0.24 -0.19 1.00
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Status distribution by sex
sex
F
M
censored
dead
status
status
censored
dead


Status distribution across health metrics
ph.ecog
3
2
1
0
censored
dead
status
ph.karno
100
90
80
70
60
50
censored
dead
status
pat.karno
100
90
80
70
60
50
30
censored
dead
status

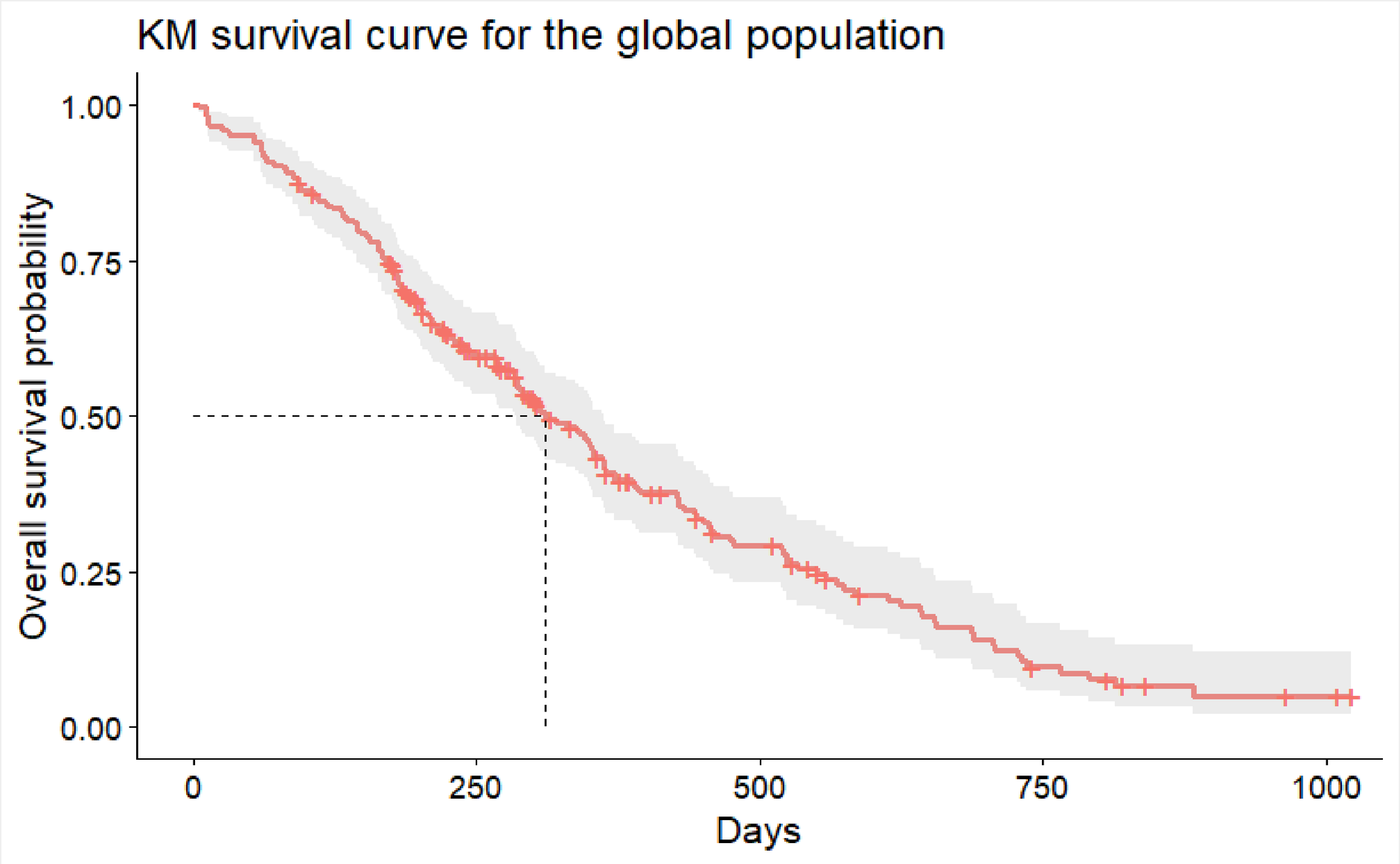

Time and weight distribution by status
Time
1000
750
500
250
0
censored
dead
Status
Weight Loss
50
25
0
-25
censored
dead
Status

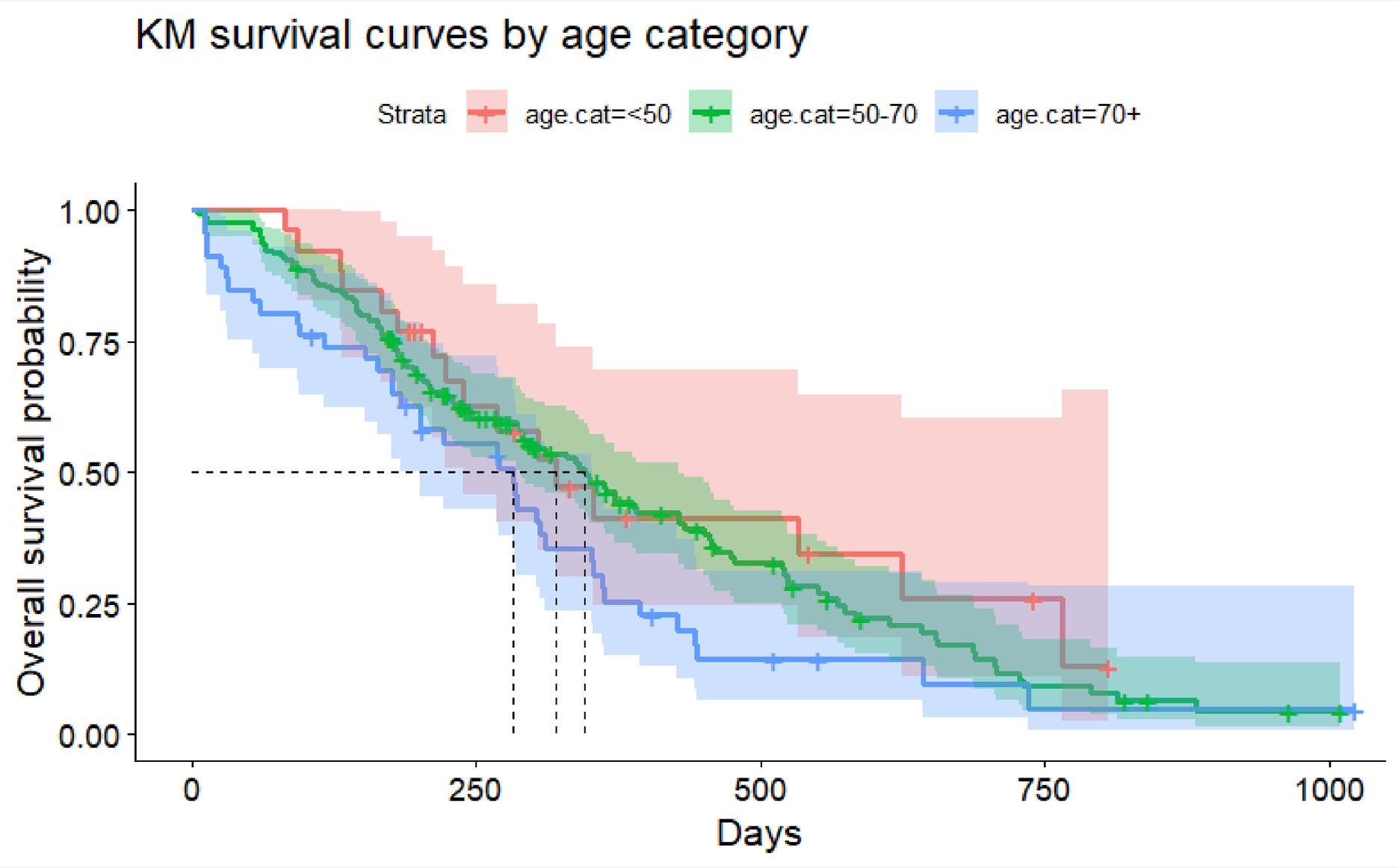

wt.loss
50
25
0
-25
0
250
500
750
1000
time

SURVIVAL ANALYSIS

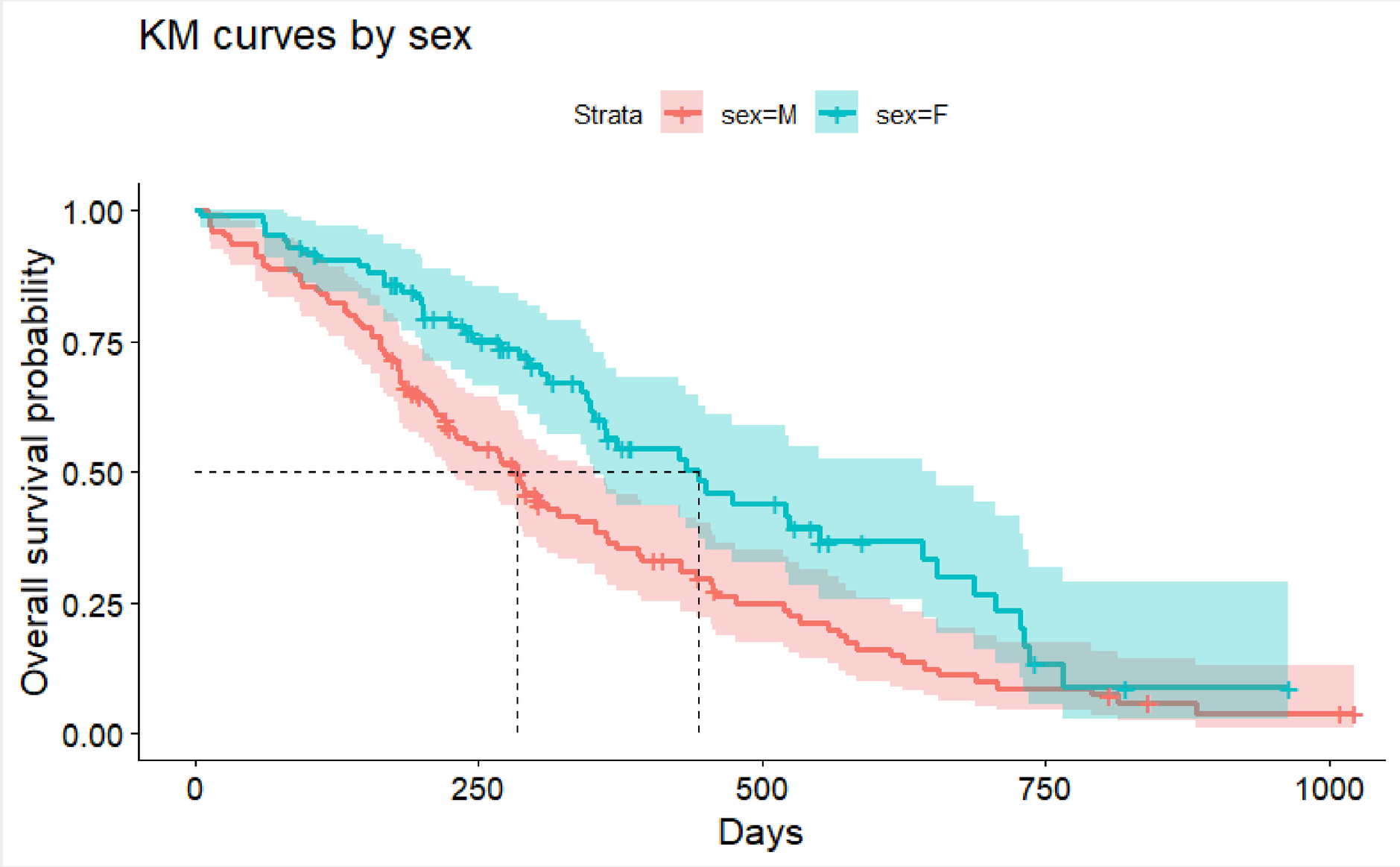
We began applying the **Kaplan–Meier estimator** to the entire dataset, as well as stratified by key categorical variables, to explore differences in survival distributions.

We then fitted several **Cox proportional hazards models** for different sets of covariates. For each model, we assessed the proportional hazards assumption, and evaluated performance through calibration and discrimination metrics.

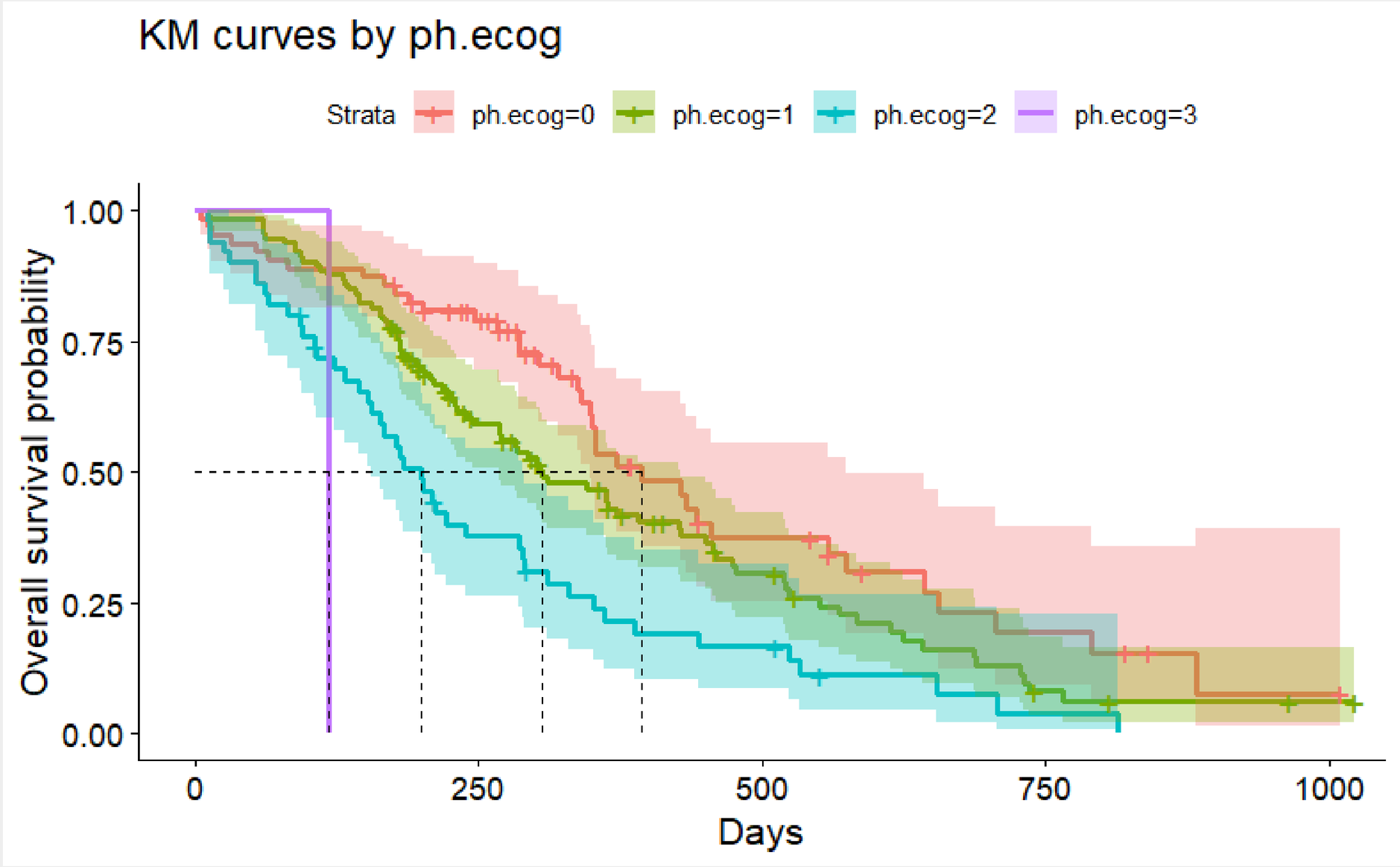

KM survival curve for the global population
Overall survival probability
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
Days



log-rank test
p= 0.08

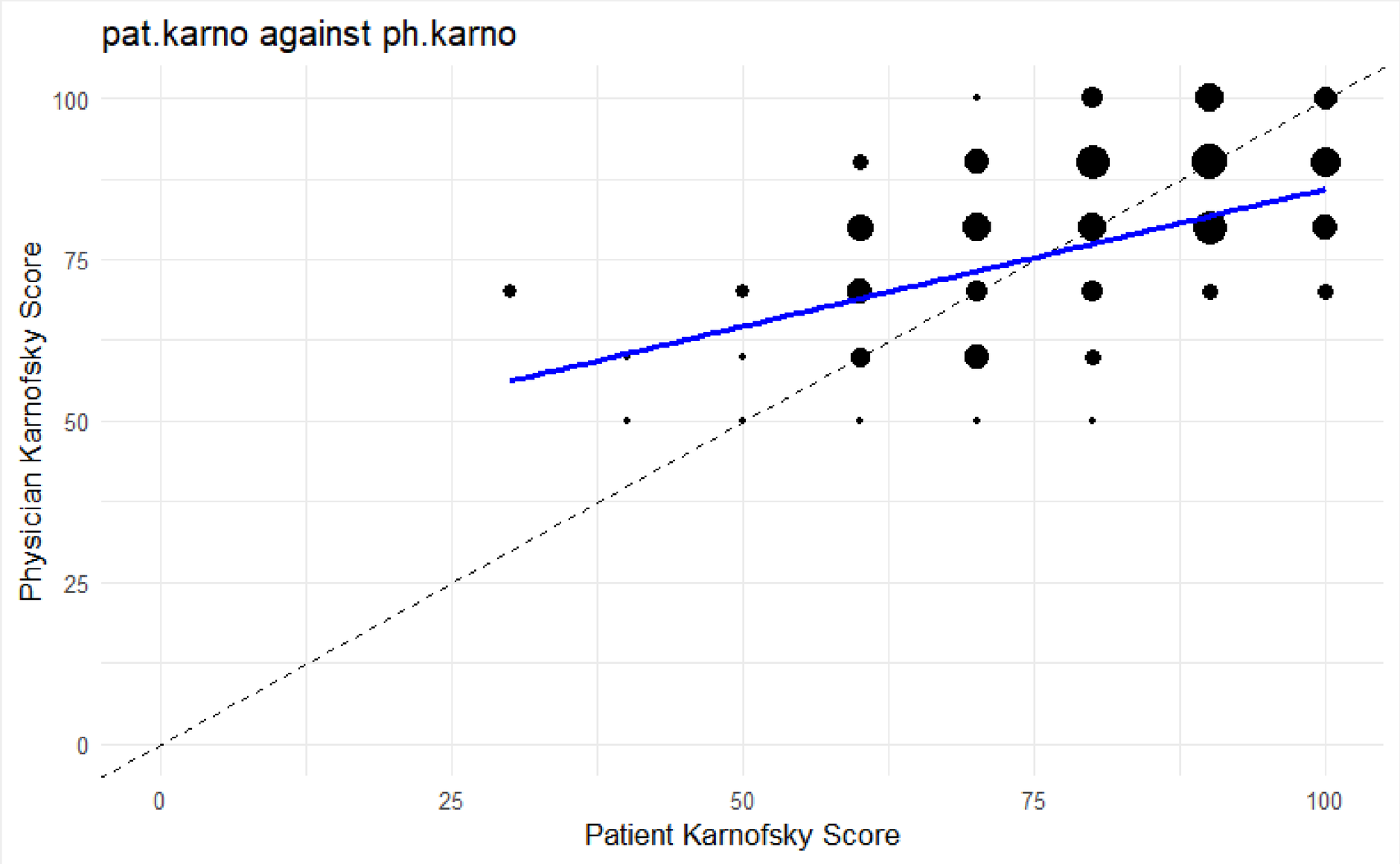


log-rank test
p= 0.002



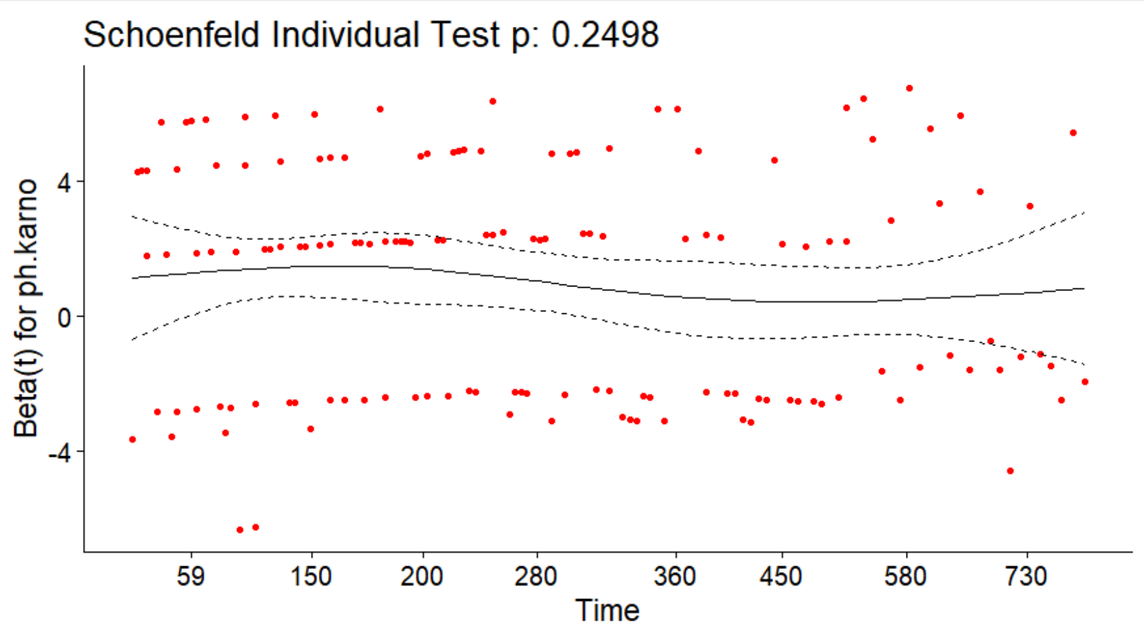
log-rank test
p= 7e-05

PATIENTS OR PHYSICIANS?


pat.karno against ph.karno
Physician Karnofsky Score
Patient Karnofsky Score
0
25
50
75
100

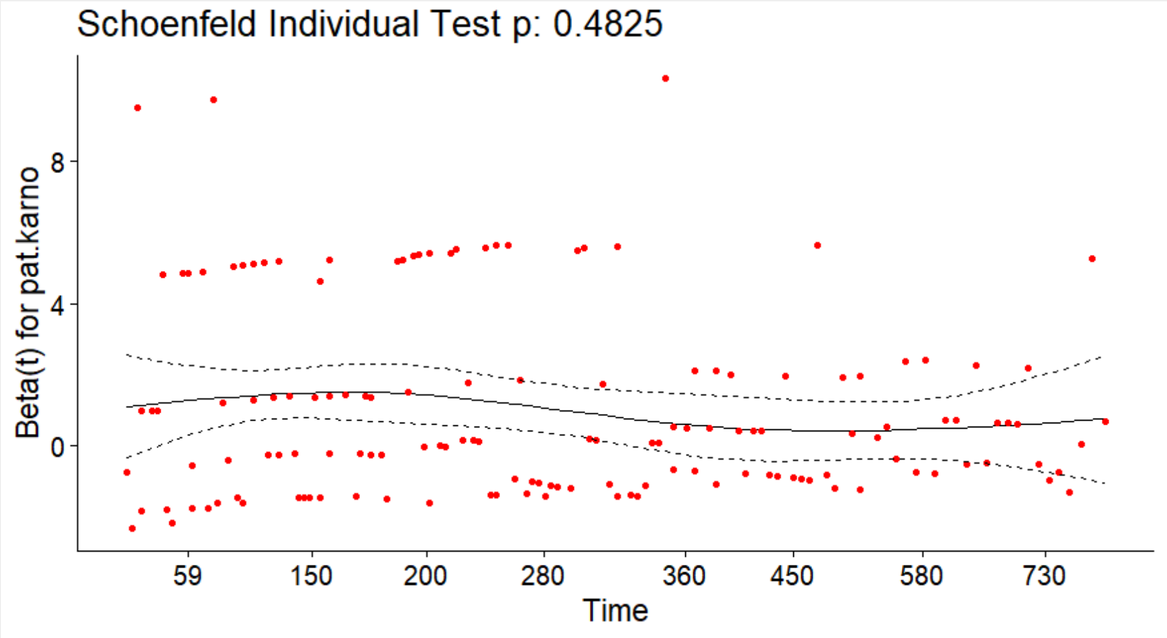
ph.karno

Concordance : 0,598
log-rank test score : 12, 91 [p = 0,02]
PH test : 6,63 [p = 0,25]
Shoenfeld Individual test :



pat.karno

Concordance : 0,610
log-rank test score : 21,23 [p = 0,003]
PH test : 6,5 [p = 0,48]
Shoenfeld Individual test :



MODELS COEFFICENTS

ph.karno

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
ph.karno60	1.0395	2.8278	0.6471	1.606	0.108
ph.karno70	0.9197	2.5086	0.6144	1.497	0.134
ph.karno80	0.7025	2.0188	0.6063	1.159	0.247
ph.karno90	0.3086	1.3615	0.6035	0.511	0.609
ph.karno100	0.2442	1.2765	0.6291	0.388	0.698

pat.karno

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
pat.karno40	-0.79712	0.45063	1.41943	-0.562	0.574
pat.karno50	0.72197	2.05849	1.15910	0.623	0.533
pat.karno60	0.07698	1.08002	1.02290	0.075	0.940
pat.karno70	-0.43726	0.64580	1.02336	-0.427	0.669
pat.karno80	-0.65501	0.51944	1.01991	-0.642	0.521
pat.karno90	-0.81713	0.44170	1.01914	-0.802	0.423
pat.karno100	-0.86409	0 42144	1.03257	-0.837	0.403

COX MODEL WITH ALL VARIABLES

Concordance : 0,677

log-rank test score : 50,01 [p = 8e-5]

	chisq	df	p
age	0.0712	1	0.79
sex	2.0663	1	0.15
ph.ecog	5.9236	3	0.12
ph.karno	5.2882	5	0.38
pat.karno	7.7598	7	0.35
wt.loss	0.1956	1	0.66
GLOBAL	18.1848	18	0.44

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
age	0.013482	1.013573	0.010595	1.273	0.203181	
sexF	-0.667947	0.512760	0.186053	-3.590	0.000331	***
ph.ecog1	0.512835	1.670019	0.302237	1.697	0.089735	.
ph.ecog2	1.071727	2.920418	0.469621	2.282	0.022483	*
ph.ecog3	2.513362	12.346369	1.155858	2.174	0.029671	*
ph.karno60	0.983394	2.673514	0.680967	1.444	0.148707	
ph.karno70	1.146855	3.148275	0.642684	1.784	0.074346	.
ph.karno80	1.460797	4.309393	0.649215	2.250	0.024443	*
ph.karno90	1.291473	3.638140	0.663390	1.947	0.051562	.
ph.karno100	1.375546	3.957238	0.728560	1.888	0.059021	.
pat.karno40	-0.115353	0.891052	1.503198	-0.077	0.938832	
pat.karno50	0.915764	2.498683	1.209555	0.757	0.448985	
pat.karno60	0.082667	1.086180	1.037507	0.080	0.936493	
pat.karno70	-0.184293	0.831692	1.058322	-0 174	0 861758	
pat.karno80	-0.268663	0.764401	1.062494	-0.253	0 800376	
pat.karno90	-0.377064	0.685872	1.066169	-0 354	0 723592	
pat.karno100	-0.515366	0.597282	1.079110	-0.478	0.632946	
wt.loss	-0.014145	0.985954	0.007465	-1.895	0.058097	.

MODEL ASSESSMENT

ROCS & AUCS

In this phase we evaluated the model using time-dependent ROC curve estimation. As time-stamps we chose [250, 365, 500, 1000]

pat.karno

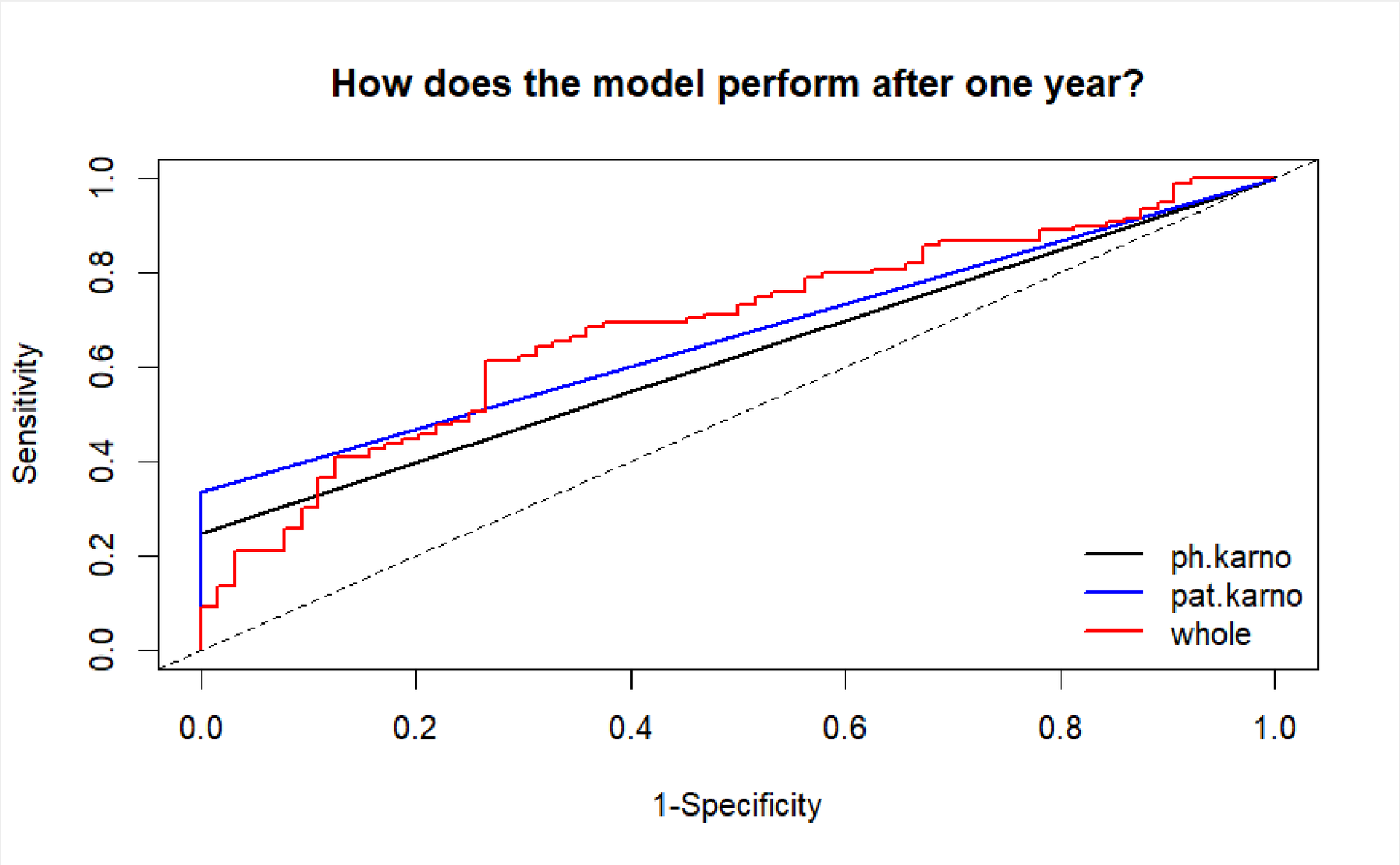
	Cases	Survivors	Censored	AUC (%)	se
t=250	75	111	23	75.16	3.94
t=365	104	64	41	66.80	4.29
t=500	120	41	48	63.82	4.92
t=1000	147	2	60	59.92	9.67

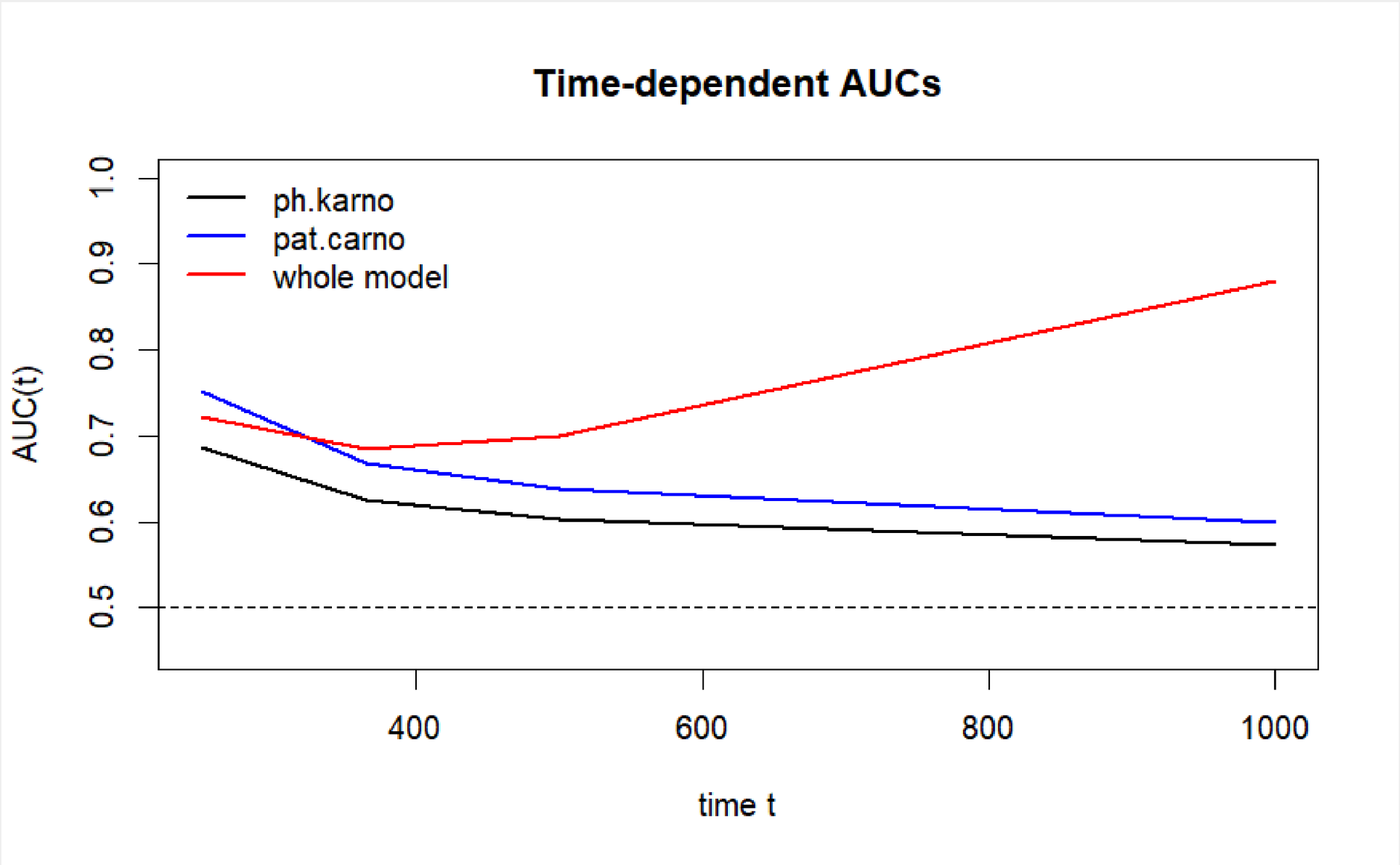
ph.karno

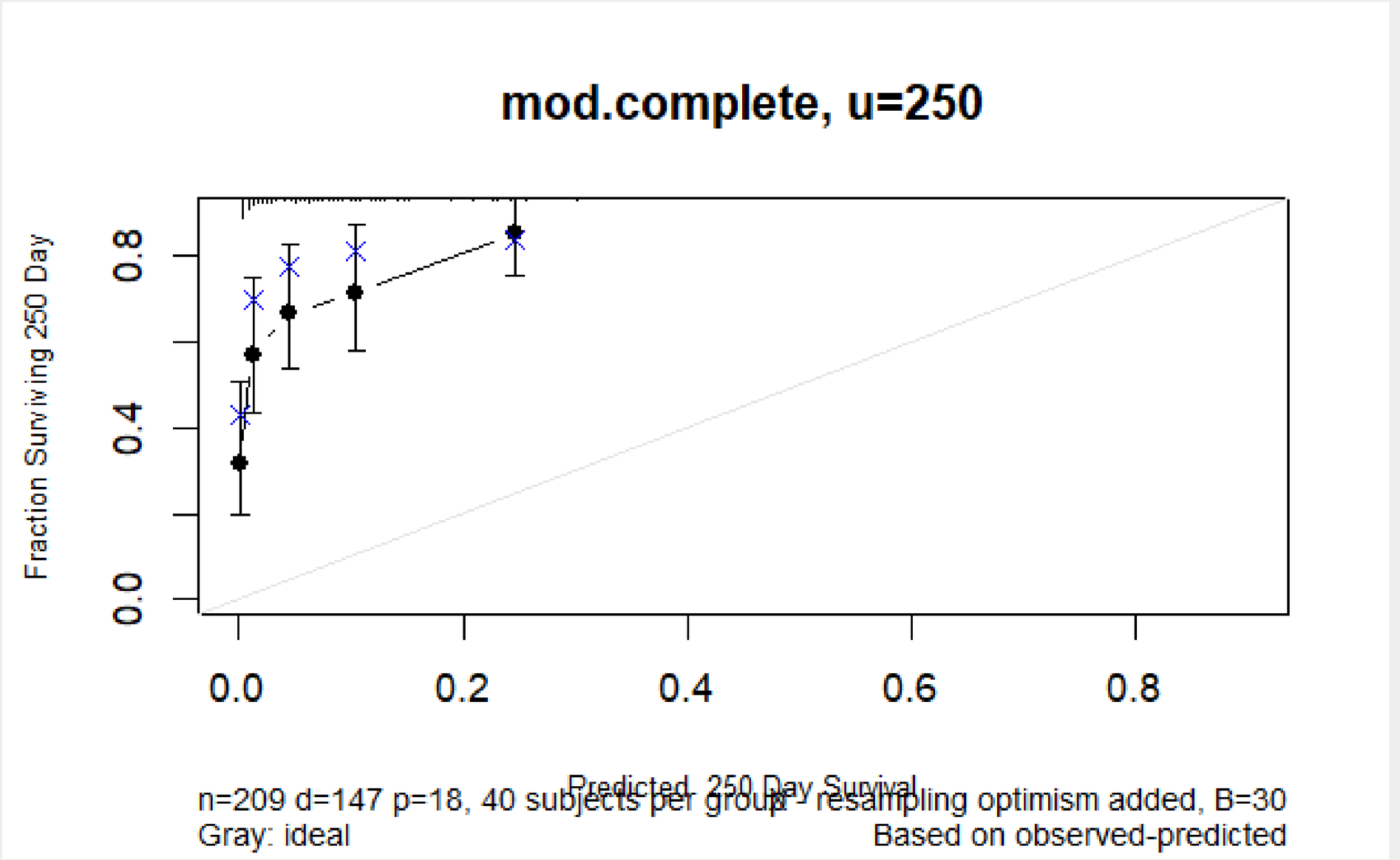
	Cases	Survivors	Censored	AUC (%)	se
t=250	75	111	23	64.13	3.90
t=365	104	64	41	59.44	4.40
t=500	120	41	48	57.76	5.34
t=1000	147	2	60	55.57	24.74

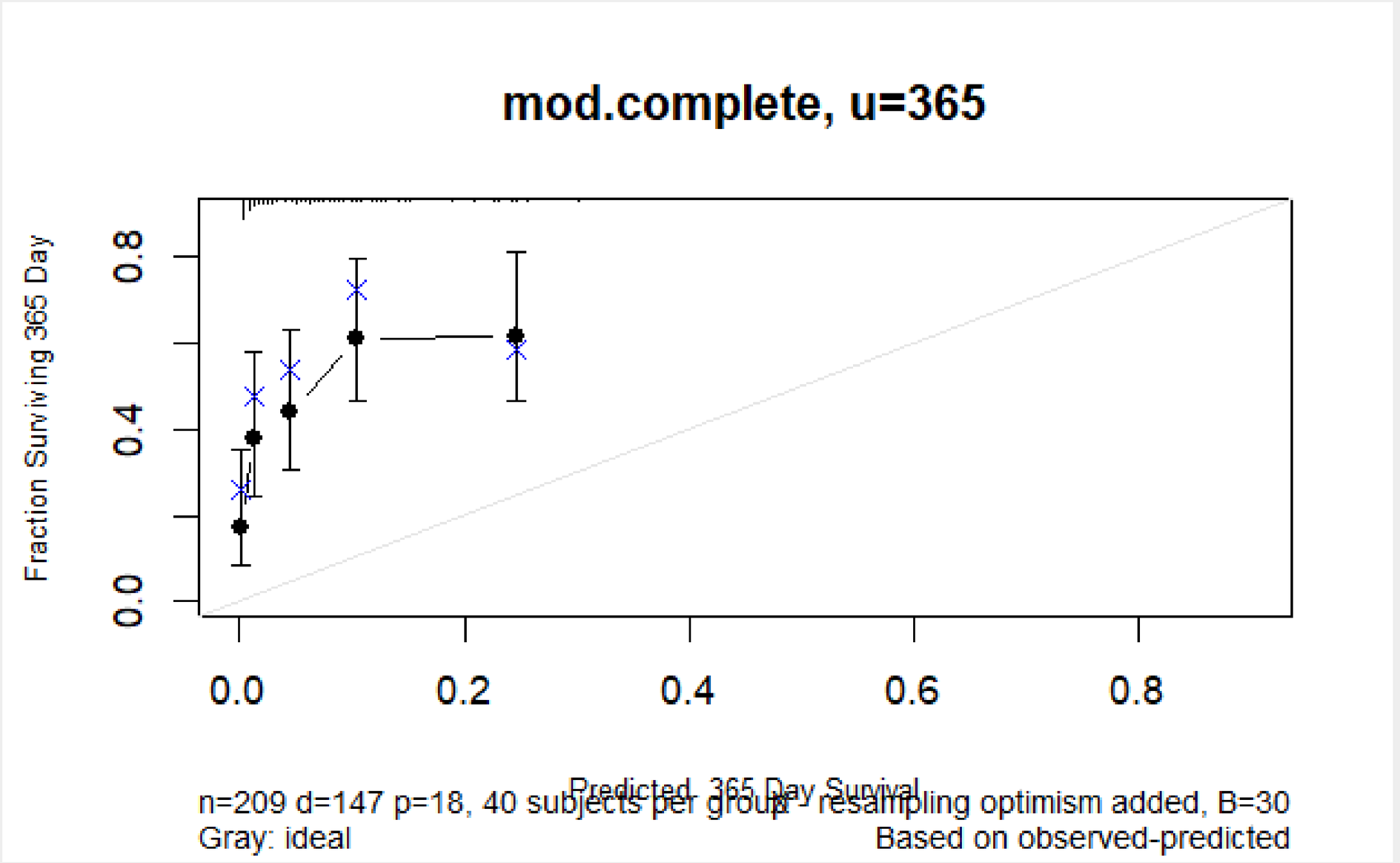
whole model

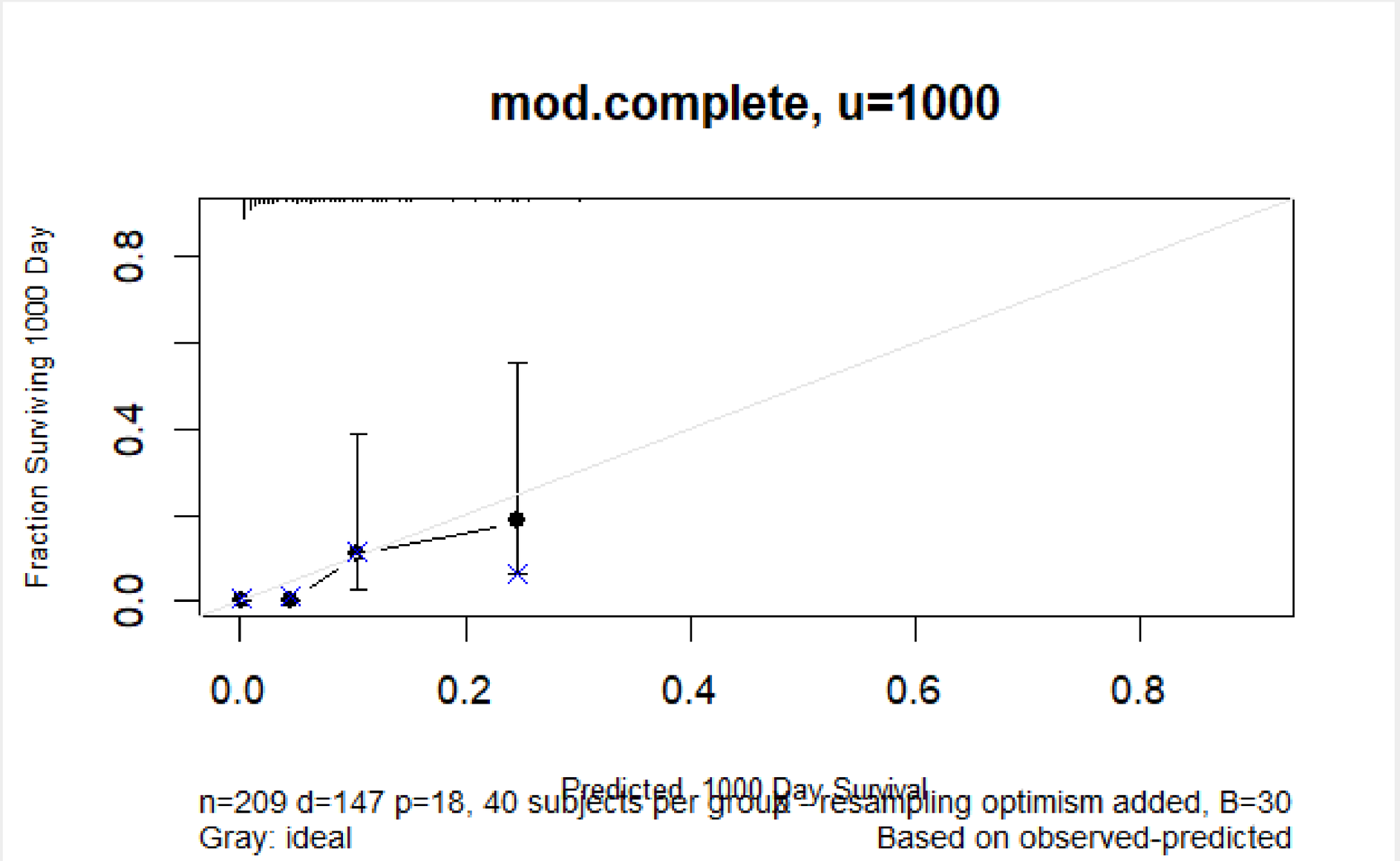
	Cases	Survivors	Censored	AUC (%)	se
t=250	75	111	23	72.16	3.77
t=365	104	64	41	68.54	4.16
t=500	120	41	48	69.96	4.57
t=1000	147	2	60	88.06	3.56

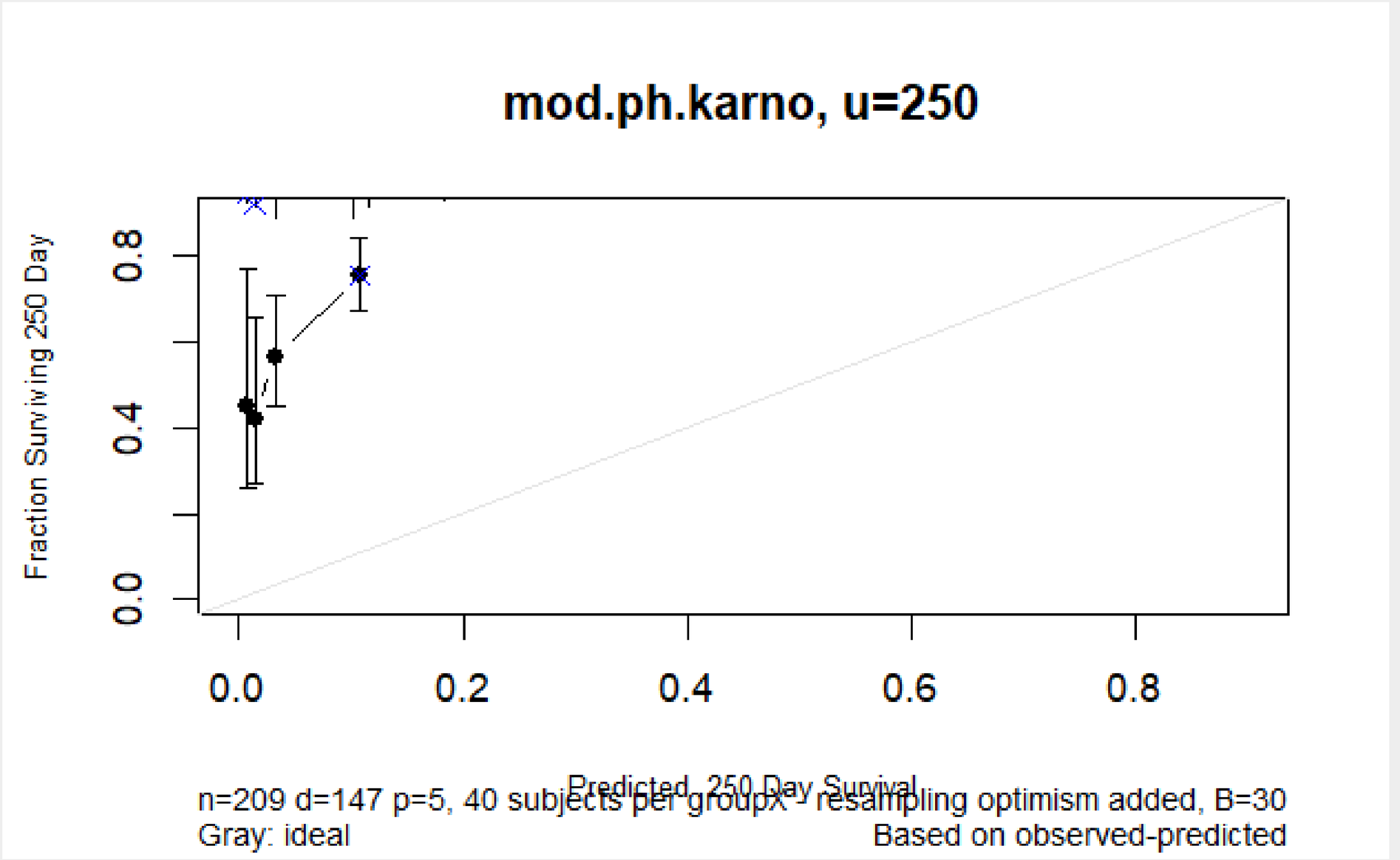

How does the model perform after one year?
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity
ph.karno
pat.karno
whole

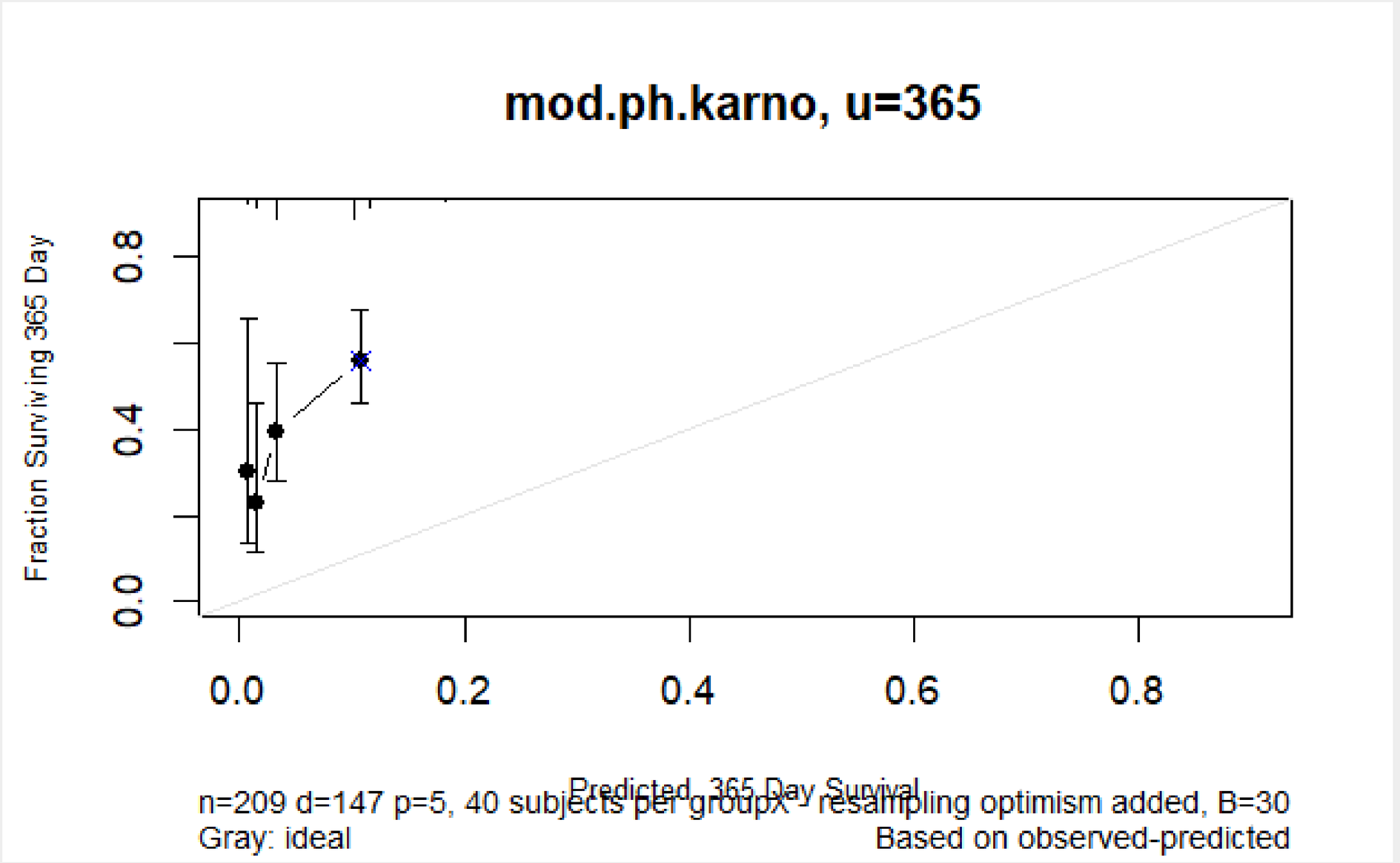

Time-dependent AUCs
ph.karno
pat.carno
whole model
AUC(t)
0.5
0.6
0.7
0.8
0.9
1.0
400
600
800
1000
time t

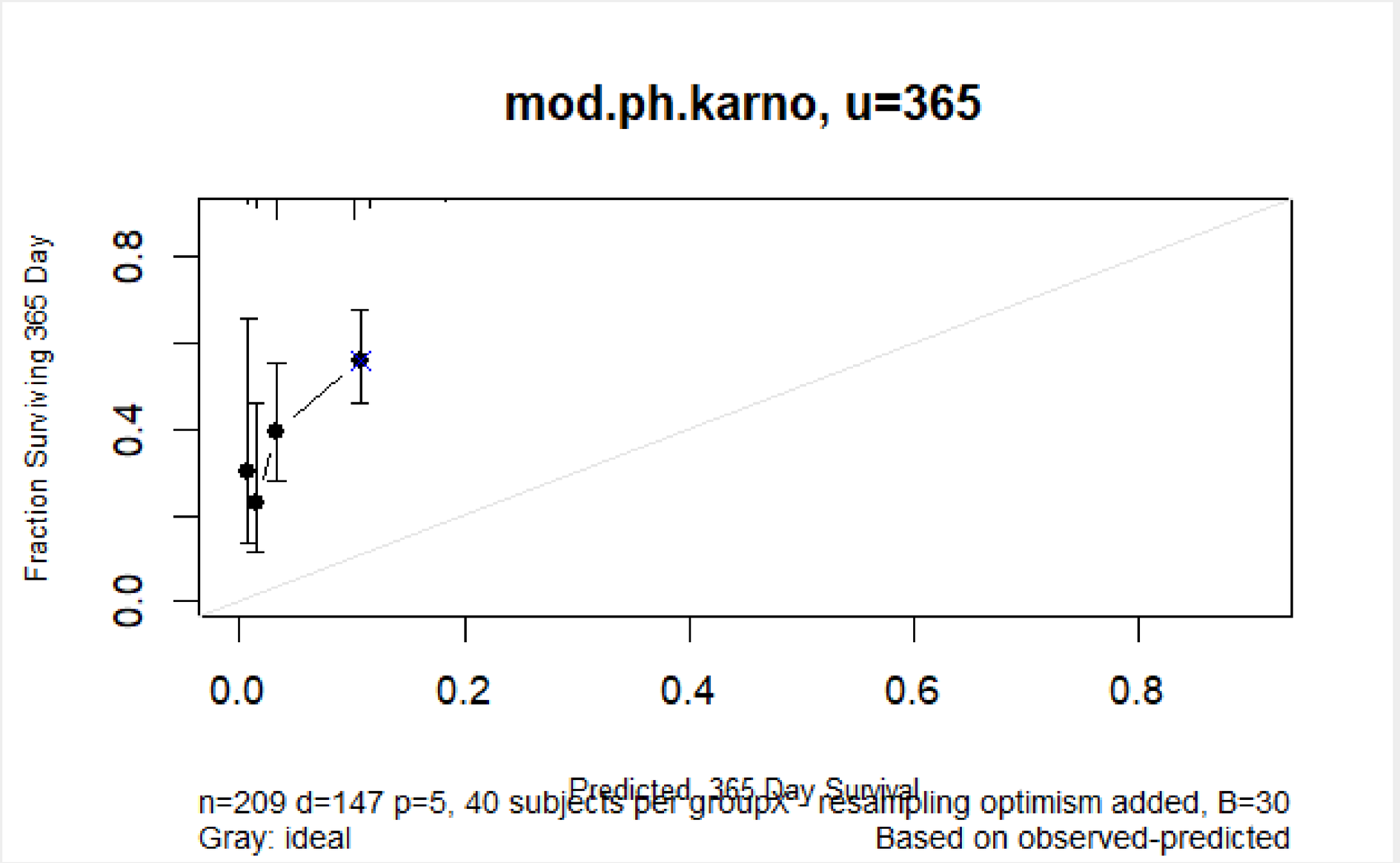

mod.complete, u=250
Fraction Surviving 250 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 250 Day Survival
n=209 d=147 p=18, 40 subjects per group; resampling optimism added, B=30
Gray: ideal
Based on observed-predicted

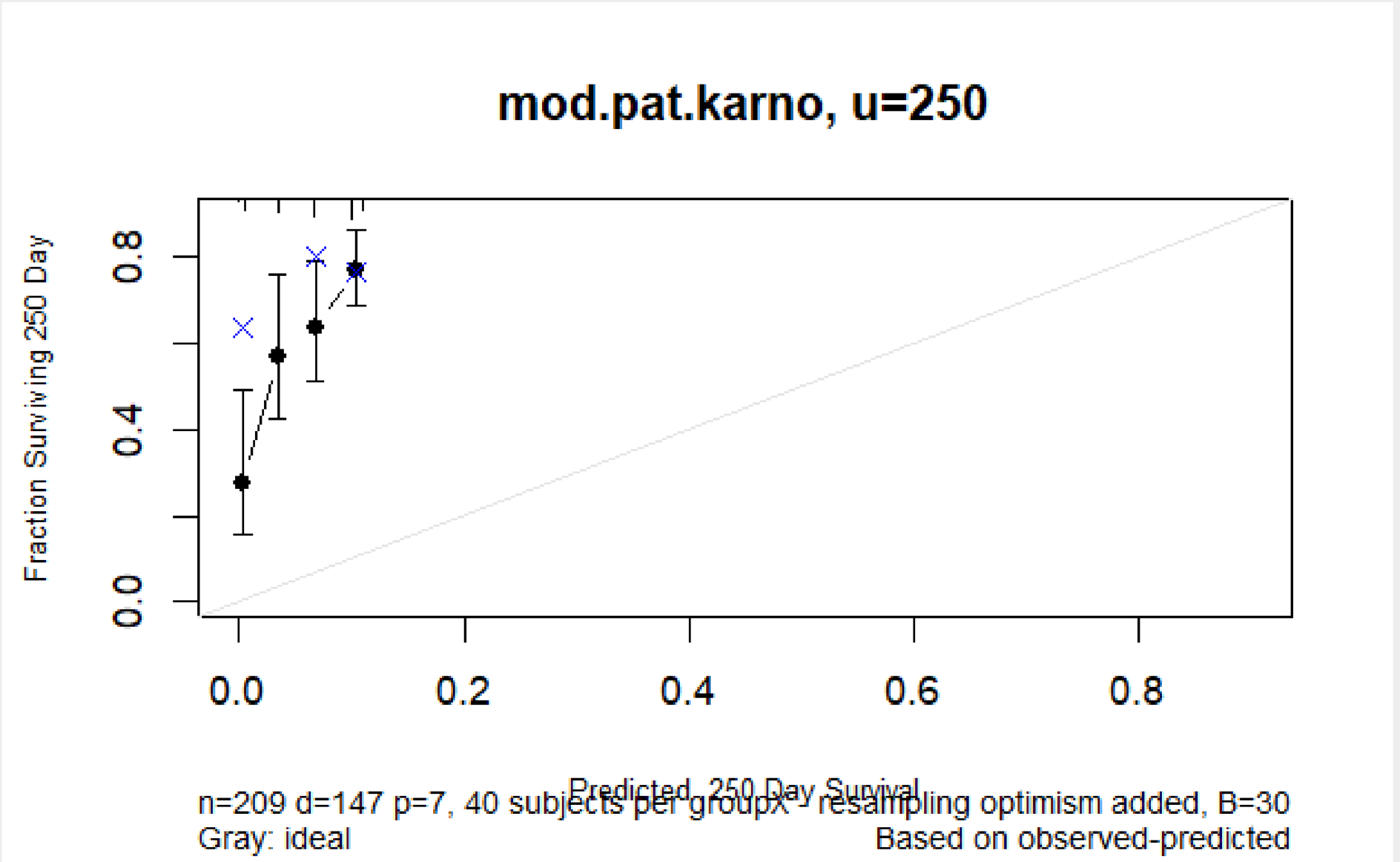

mod.complete, u=365
Fraction Surviving 365 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 365 Day Survival
n=209 d=147 p=18, 40 subjects per group; resampling optimism added, B=30
Gray: ideal
Based on observed-predicted

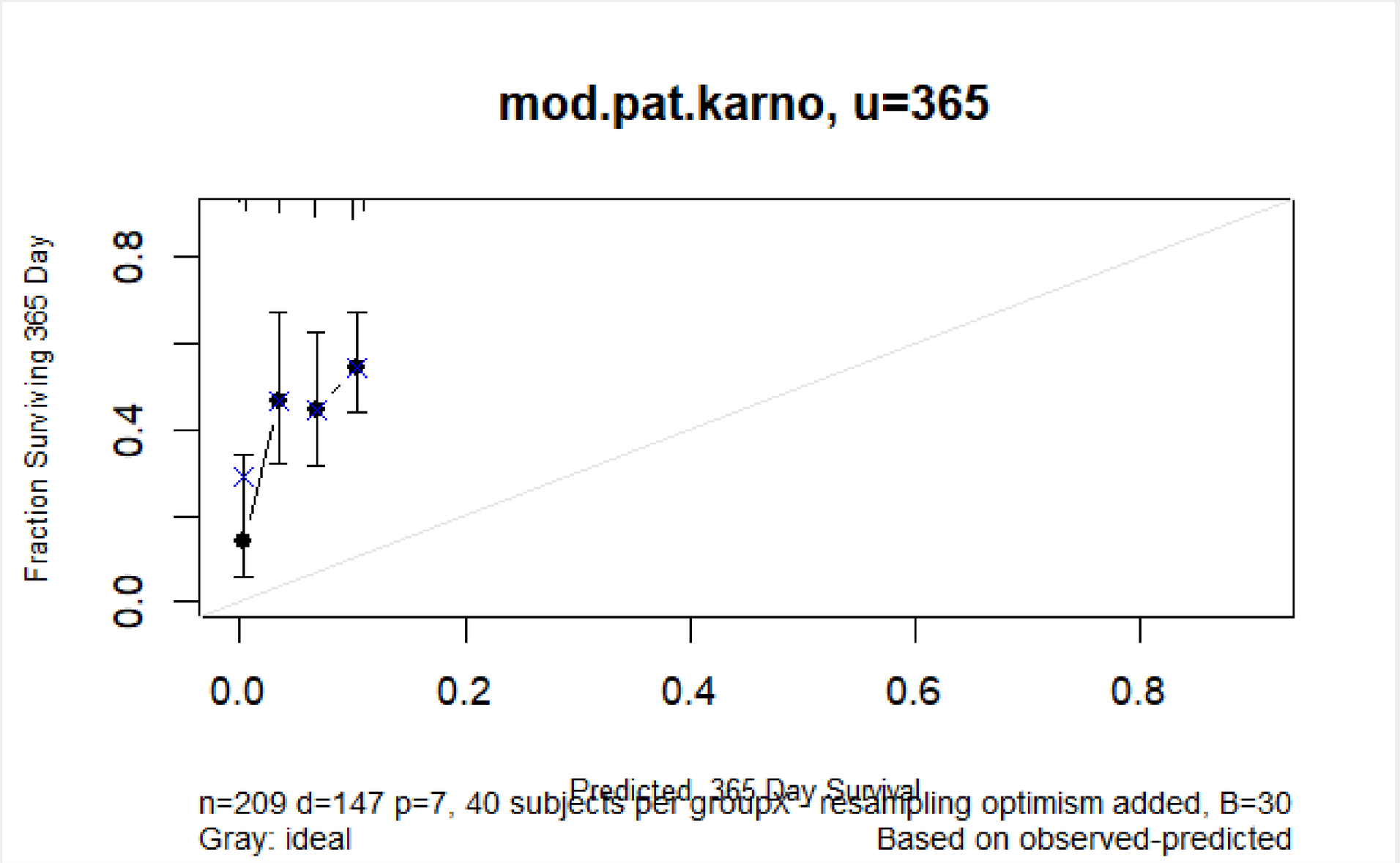

mod.complete, u=1000
Fraction Surviving 1000 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 1000 Day Survival
n=209 d=147 p=18, 40 subjects per group resampling optimism added, B=30
Gray: ideal
Based on observed-predicted

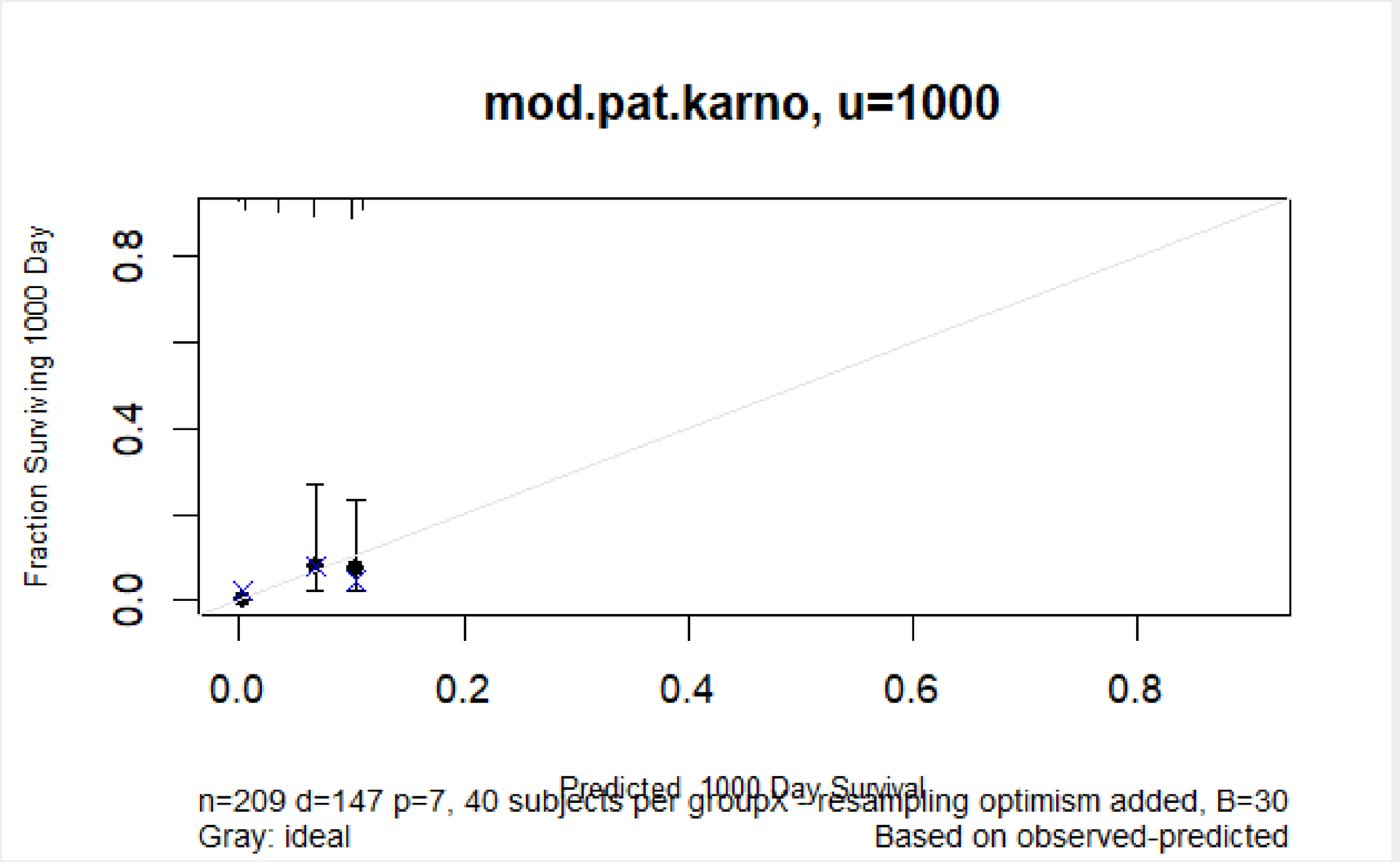

mod.ph.karno, u=250
Fraction Surviving 250 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 250 Day Survival
n=209 d=147 p=5, 40 subjects per group; resampling optimism added, B=30
Gray: ideal
Based on observed-predicted


mod.ph.karno, u=365
Fraction Surviving 365 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 365 Day Survival
n=209 d=147 p=5, 40 subjects per group; resampling optimism added, B=30
Gray: ideal
Based on observed-predicted


mod.ph.karno, u=365
Fraction Surviving 365 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 365 Day Survival
n=209 d=147 p=5, 40 subjects per group; resampling optimism added, B=30
Gray: ideal
Based on observed-predicted


mod.pat.karno, u=250
Fraction Surviving 250 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 250 Day Survival
n=209 d=147 p=7, 40 subjects per group; resampling optimism added, B=30
Gray: ideal
Based on observed-predicted


mod.pat.karno, u=365
Fraction Surviving 365 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 365 Day Survival
n=209 d=147 p=7, 40 subjects per groupX - resampling optimism added, B=30
Gray: ideal
Based on observed-predicted


mod.pat.karno, u=1000
Fraction Surviving 1000 Day
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
Predicted 1000 Day Survival
n=209 d=147 p=7, 40 subjects per group
x - resampling optimism added, B=30
Gray: ideal
Based on observed-predicted

A NOTE ABOUT VALIDATION

We tried to perform internal validation for our analysis but we found that in order to have a reliable estimate we would have needed to sacrifice too much of the observations, resulting in a less reliable model, and test.

We consider the estimates on the training data as an optimistic estimate at best.