Chapter 21
Case Study in Cox Regression

21.1 Choosing the Number of Parameters and Fitting the Model

Consider the randomized trial of estrogen for treatment of prostate cancer[87] described in Chapter 8. Let us now develop a model for time until death (of any cause). There are 354 deaths among the 502 patients. To be able to efficiently estimate treatment benefit, to test for differential treatment effect, or to estimate prognosis or absolute treatment benefit for individual patients, we need a multivariable survival model. In this case study we do not make use of data reductions obtained in Chapter 8 but show simpler (partial) approaches to data reduction. We do use the `transcan` results for imputation.

First let's assess the wisdom of fitting a full additive model that does not assume linearity of effect for any predictor. Categorical predictors are expanded using dummy variables. For `pf` we could lump the last two categories as before since the last category has only two patients. Likewise, we could combine the last two levels of `ekg`. Continuous predictors are expanded by fitting four-knot restricted cubic spline functions, which contain two nonlinear terms and thus have a total of three d.f. Table 21.1 defines the candidate predictors and lists their d.f. The variable `stage` is not listed as it can be predicted with high accuracy from `sz,sg,ap,bm` (`stage` could have been used as a predictor for imputing missing values on `sz, sg`). There are a total of 36 candidate d.f. that should not be artificially reduced by "univariable screening" or graphical assessments of association with death. This is about 1/10 as many predictor d.f. as there are deaths, so there is some hope that a fitted model may validate. Let us also examine this issue by estimating the amount of shrinkage using Equation 4.3. We first use `transcan` impute missing data.

```
require(rms)
```


F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_21

Table 21.1 Initial allocation of degrees of freedom

Predictor	Name	d.f.	Original Levels
Dose of estrogen	`rx`	3	placebo, 0.2, 1.0, 5.0 mg estrogen
Age in years	`age`	3	
Weight index: wt(kg)−ht(cm)+200	`wt`	3	
Performance rating	`pf`	2	normal, in bed $< 50\%$ of time, in bed $> 50\%$, in bed always
History of cardiovascular disease	`hx`	1	present/absent
Systolic blood pressure/10	`sbp`	3	
Diastolic blood pressure/10	`dbp`	3	
Electrocardiogram code	`ekg`	5	normal, benign, rhythm disturb., block, strain, old myocardial infarction, new MI
Serum hemoglobin (g/100ml)	`hg`	3	
Tumor size (cm^2)	`sz`	3	
Stage/histologic grade combination	`sg`	3	
Serum prostatic acid phosphatase	`ap`	3	
Bone metastasis	`bm`	1	present/absent

```
getHdata(prostate)
levels(prostate$ekg)[levels(prostate$ekg) %in%
                     c('old MI','recent MI')] ← 'MI'
# combines last 2 levels and uses a new name, MI

prostate$pf.coded ← as.integer(prostate$pf)
# save original pf, re-code to 1-4
levels(prostate$pf)  ← c(levels(prostate$pf)[1:3],
                         levels(prostate$pf)[3])
# combine last 2 levels

w ← transcan(∼ sz + sg + ap + sbp + dbp + age +
             wt + hg + ekg + pf + bm + hx, imputed=TRUE,
             data=prostate, pl=FALSE, pr=FALSE)

attach(prostate)
sz  ← impute(w, sz, data=prostate)
sg  ← impute(w, sg, data=prostate)
age ← impute(w, age,data=prostate)
wt  ← impute(w, wt, data=prostate)
ekg ← impute(w, ekg,data=prostate)

dd ← datadist(prostate); options(datadist='dd')
```

```
units(dtime) ← 'Month'
S ← Surv(dtime, status != 'alive')

f ← cph(S ∼ rx + rcs(age,4) + rcs(wt,4) + pf + hx +
         rcs(sbp,4) + rcs(dbp,4) + ekg + rcs(hg,4) +
         rcs(sg,4) + rcs(sz,4) + rcs(log(ap),4) + bm)
```

```
print(f, latex=TRUE, coefs=FALSE)
```

Cox Proportional Hazards Model

```
cph(formula = S ~ rx + rcs(age, 4) + rcs(wt, 4) + pf + hx
    + rcs(sbp, 4) + rcs(dbp, 4) + ekg + rcs(hg, 4)
    + rcs(sg, 4) + rcs(sz, 4) + rcs(log(ap), 4) + bm)
```

		Model Tests		Discrimination Indexes	
Obs	502	LR χ^2	136.22	R^2	0.238
Events	354	d.f.	36	D_{xy}	0.333
Center	-2.9933	Pr($> \chi^2$)	0.0000	g	0.787
		Score χ^2	143.62	g_r	2.196
		Pr($> \chi^2$)	0.0000		

The likelihood ratio χ^2 statistic is 136.2 with 36 d.f. This test is highly significant so some modeling is warranted. The AIC value (on the χ^2 scale) is $136.2-2\times36 = 64.2$. The rough shrinkage estimate is 0.74 (100.2/136.2) so we estimate that 0.26 of the model fitting will be noise, especially with regard to calibration accuracy. The approach of Spiegelhalter[582] is to fit this full model and to shrink predicted values. We instead try to do data reduction (blinded to individual χ^2 statistics from the above model fit) to see if a reliable model can be obtained without shrinkage. A good approach at this point might be to do a variable clustering analysis followed by single degree of freedom scoring for individual predictors or for clusters of predictors. Instead we do an informal data reduction. The strategy is described in Table 21.2. For `ap`, more exploration is desired to be able to model the shape of effect with such a highly skewed distribution. Since we expect the tumor variables to be strong prognostic factors we retain them as separate variables. No assumption is made for the dose-response shape for estrogen, as there is reason to expect a non-monotonic effect due to competing risks for cardiovascular death.

```
heart ← hx + ekg %nin% c('normal','benign')
label(heart) ← 'Heart Disease Code'
map   ← (2*dbp + sbp)/3
label(map) ← 'Mean Arterial Pressure/10'
dd ← datadist(dd, heart, map)

f ← cph(S ∼ rx + rcs(age,4) + rcs(wt,3) + pf.coded +
```

Table 21.2 Final allocation of degrees of freedom

Variables	Reductions	d.f. Saved
`wt`	Assume variable not important enough for 4 knots; use 3 knots	1
`pf`	Assume linearity	1
`hx,ekg`	Make new 0,1,2 variable and assume linearity: 2 = `hx` and `ekg` not normal or benign, 1 = either, 0 = none	5
`sbp,dbp`	Combine into mean arterial bp and use 3 knots: map = (2 `dbp` + `sbp`)/3	4
`sg`	Use 3 knots	1
`sz`	Use 3 knots	1
`ap`	Look at shape of effect of `ap` in detail, and take log before expanding as spline to achieve numerical stability: add 1 knots	−1

```
        heart + rcs(map,3) + rcs(hg,4) +
        rcs(sg,3) + rcs(sz,3) + rcs(log(ap),5) + bm,
        x=TRUE, y=TRUE, surv=TRUE, time.inc=5*12)
print(f, latex=TRUE, coefs=3)
```

Cox Proportional Hazards Model

```
cph(formula = S ~ rx + rcs(age, 4) + rcs(wt, 3) + pf.coded +
    heart + rcs(map, 3) + rcs(hg, 4) + rcs(sg, 3) +
    rcs(sz, 3) + rcs(log(ap), 5) + bm, x = TRUE, y = TRUE,
    surv = TRUE, time.inc = 5 * 12)
```

		Model Tests		Discrimination Indexes	
Obs	502	LR χ^2	118.37	R^2	0.210
Events	354	d.f.	24	D_{xy}	0.321
Center	-2.4307	Pr($>\chi^2$)	0.0000	g	0.717
		Score χ^2	125.58	g_r	2.049
		Pr($>\chi^2$)	0.0000		

	Coef	S.E.	Wald Z	Pr($>\|Z\|$)
rx=0.2 mg estrogen	-0.0002	0.1493	0.00	0.9987
rx=1.0 mg estrogen	-0.4160	0.1657	-2.51	0.0121
rx=5.0 mg estrogen	-0.1107	0.1571	-0.70	0.4812
...				

Table 21.3 Wald Statistics for `S`

	χ^2	d.f.	P
rx	8.01	3	0.0459
age	13.84	3	0.0031
Nonlinear	9.06	2	0.0108
wt	8.21	2	0.0165
Nonlinear	2.54	1	0.1110
pf.coded	3.79	1	0.0517
heart	23.51	1	< 0.0001
map	0.04	2	0.9779
Nonlinear	0.04	1	0.8345
hg	12.52	3	0.0058
Nonlinear	8.25	2	0.0162
sg	1.64	2	0.4406
Nonlinear	0.05	1	0.8304
sz	12.73	2	0.0017
Nonlinear	0.06	1	0.7990
ap	6.51	4	0.1639
Nonlinear	6.22	3	0.1012
bm	0.03	1	0.8670
TOTAL NONLINEAR	23.81	11	0.0136
TOTAL	119.09	24	< 0.0001

```
# x, y for predict, validate, calibrate;
# surv, time.inc for calibrate
latex(anova(f),file='',label='tab:coxcase-anova1')# Table 21.3
```

The total savings is thus 12 d.f. The likelihood ratio χ^2 is 118 with 24 d.f., with a slightly improved AIC of 70. The rough shrinkage estimate is slightly better at 0.80, but still worrisome. A further data reduction could be done, such as using the `transcan` transformations determined from self-consistency of predictors, but we stop here and use this model.

From Table 21.3 there are 11 parameters associated with nonlinear effects, and the overall test of linearity indicates the strong presence of nonlinearity for at least one of the variables `age,wt,map,hg,sz,sg,ap`. There is no strong evidence for a difference in survival time between doses of estrogen.

21.2 Checking Proportional Hazards

Now that we have a tentative model, let us examine the model's distributional assumptions using smoothed scaled Schoenfeld residuals. A messy detail is how to handle multiple regression coefficients per predictor. Here we do an

approximate analysis in which each predictor is scored by adding up all that predictor's terms in the model, to transform that predictor to optimally relate to the log hazard (at least if the *shape* of the effect does not change with time). In doing this we are temporarily ignoring the fact that the individual regression coefficients were estimated from the data. For dose of estrogen, for example, we code the effect as 0 (placebo), −0.00025 (0.2 mg), −0.416 (1.0 mg), and −0.111 (5.0 mg), and `age` is transformed using its fitted spline function. In the `rms` package the `predict` function easily summarizes multiple terms and produces a matrix (here, `z`) containing the total effects for each predictor. Matrix factors can easily be included in model formulas.

```
z ← predict(f, type='terms')
# required x=T above to store design matrix
f.short ← cph(S ~ z, x=TRUE, y=TRUE)
# store raw x, y so can get residuals
```

The fit `f.short` based on the matrix of single d.f. predictors `z` has the same LR χ^2 of 118 as the fit `f`, but with a falsely low 11 d.f. All regression coefficients are unity.

Now we compute scaled Schoenfeld residuals separately for each predictor and test the PH assumption using the "correlation with time" test. Also plot smoothed trends in the residuals. The `plot` method for `cox.zph` objects uses cubic splines to smooth the relationship.

```
phtest ← cox.zph(f.short, transform='identity')
phtest
```

```
              rho    chisq      p
rx        0.10232 4.00823 0.0453
age      -0.05483 1.05850 0.3036
wt        0.01838 0.11632 0.7331
pf.coded -0.03429 0.41884 0.5175
heart     0.02650 0.30052 0.5836
map       0.02055 0.14135 0.7069
hg       -0.00362 0.00511 0.9430
sg       -0.05137 0.94589 0.3308
sz       -0.01554 0.08330 0.7729
ap        0.01720 0.11858 0.7306
bm        0.04957 0.95354 0.3288
GLOBAL         NA 7.18985 0.7835
```

```
plot(phtest, var='rx')    # Figure 21.1
```

Perhaps only the drug effect significantly changes over time ($P = 0.05$ for testing the correlation `rho` between the scaled Schoenfeld residual and time), but when a global test of PH is done penalizing for 11 d.f., the P value is 0.78. A graphical examination of the trends doesn't find anything interesting for the last 10 variables. A residual plot is drawn for `rx` alone and is shown in Figure 21.1. We ignore the possible increase in effect of estrogen over time. If this non-PH is real, a more accurate model might be obtained by stratifying on `rx` or by using a time × `rx` interaction as a time-dependent covariable.

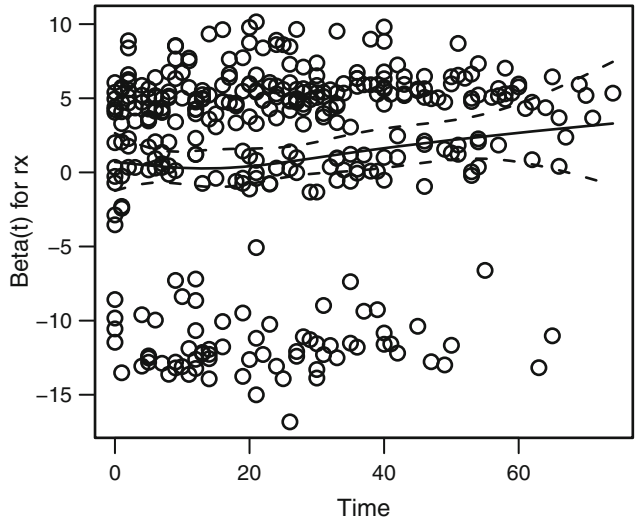

Fig. 21.1 Raw and spline-smoothed scaled Schoenfeld residuals for dose of estrogen, nonlinearly coded from the Cox model fit, with ± 2 standard errors.

21.3 Testing Interactions

Note that the model has several insignificant predictors. These are not deleted, as that would not improve predictive accuracy and it would make accurate confidence intervals hard to obtain. At this point it would be reasonable to test prespecified interactions. Here we test all interactions with dose. Since the multiple terms for many of the predictors (and for `rx`) make for a great number of d.f. for testing interaction (and a loss of power), we do approximate tests on the data-driven coding of predictors. *P*-values for these tests are likely to be somewhat anti-conservative.

```
z.dose ← z[,"rx"]  # same as saying z[,1] - get first column
z.other ← z[,-1]   # all but the first column of z
f.ia ← cph(S ∼ z.dose * z.other)   # Figure 21.4:
latex(anova(f.ia), file='', label='tab:coxcase-anova2')
```

The global test of additivity in Table 21.4 has $P = 0.27$, so we ignore the interactions (and also forget to penalize for having looked for them below!).

21.4 Describing Predictor Effects

Let us plot how each predictor is related to the log hazard of death, including 0.95 confidence bands. Note in Figure 21.2 that due to a peculiarity of the Cox model the standard error of the predicted $X\hat{\beta}$ is zero at the reference values (medians here, for continuous predictors).

Table 21.4 Wald Statistics for `S`

	χ^2	d.f.	P
z.dose (Factor+Higher Order Factors)	18.74	11	0.0660
All Interactions	12.17	10	0.2738
z.other (Factor+Higher Order Factors)	125.89	20	< 0.0001
All Interactions	12.17	10	0.2738
z.dose × z.other (Factor+Higher Order Factors)	12.17	10	0.2738
TOTAL	129.10	21	< 0.0001

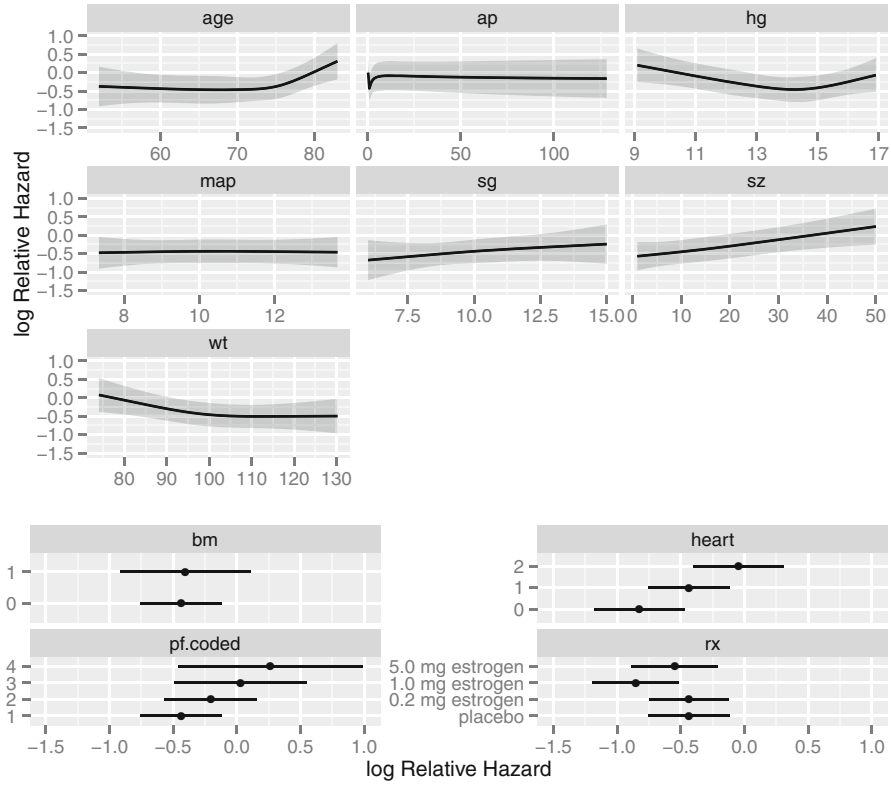

Fig. 21.2 Shape of each predictor on log hazard of death. Y-axis shows $X\hat{\beta}$, but the predictors not plotted are set to reference values. Note the highly non-monotonic relationship with **ap**, and the increased slope after age 70 which occurs in outcome models for various diseases.

```
ggplot(Predict(f), sepdiscrete='vertical', nlevels=4,
       vnames='names')  # Figure 21.2
```

21.5 Validating the Model

We first validate this model for Somers' D_{xy} rank correlation between predicted log hazard and observed survival time, and for slope shrinkage. The bootstrap is used (with 300 resamples) to penalize for possible overfitting, as discussed in Section 5.3.

```
set.seed(1)  # so can reproduce results
v ← validate(f, B=300)
```

Divergence or singularity in 83 samples

```
latex(v, file='')
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.3208	0.3454	0.2954	0.0500	0.2708	217
R^2	0.2101	0.2439	0.1754	0.0685	0.1417	217
Slope	1.0000	1.0000	0.7941	0.2059	0.7941	217
D	0.0292	0.0348	0.0238	0.0110	0.0182	217
U	−0.0005	−0.0005	0.0023	−0.0028	0.0023	217
Q	0.0297	0.0353	0.0216	0.0138	0.0159	217
g	0.7174	0.7918	0.6273	0.1645	0.5529	217

Here "training" refers to accuracy when evaluated on the bootstrap sample used to fit the model, and "test" refers to the accuracy when this model is applied without modification to the original sample. The apparent D_{xy} is 0.32, but a better estimate of how well the model will discriminate prognoses in the future is $D_{xy} = 0.27$. The bootstrap estimate of slope shrinkage is 0.79, close to the simple heuristic estimate. The shrinkage coefficient could easily be used to shrink predictions to yield better calibration.

Finally, we validate the model (without using the shrinkage coefficient) for calibration accuracy in predicting the probability of surviving five years. The bootstrap is used to estimate the optimism in how well predicted five-year survival from the final Cox model tracks flexible smooth estimates, without any binning of predicted survival probabilities or assuming proportional hazards.

```
cal ← calibrate(f, B=300, u=5*12, maxdim=4)
```

```
Using Cox survival estimates at 60 Months
```

```
plot(cal, subtitles=FALSE)    # Figure 21.3
```

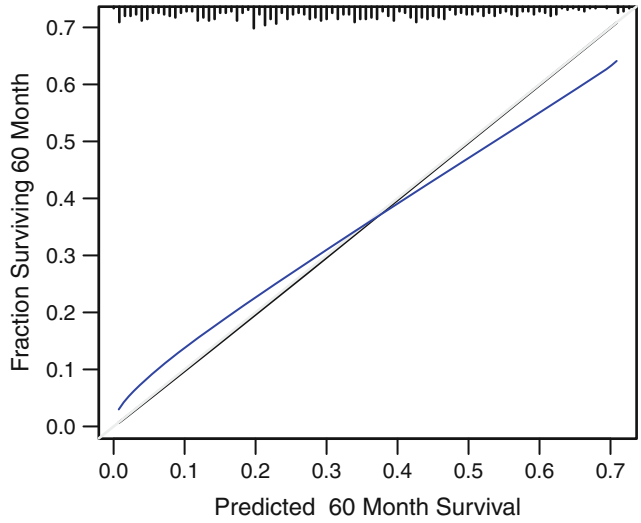

Fig. 21.3 Bootstrap estimate of calibration accuracy for 5-year estimates from the final Cox model, using adaptive linear spline hazard regression[361]. The line nearer the ideal line corresponds to apparent predictive accuracy. The blue curve corresponds to bootstrap-corrected estimates.

The estimated calibration curves are shown in Figure 21.3, similar to what was done in Figure 19.11. Bootstrap calibration demonstrates some overfitting, consistent with regression to the mean. The absolute error is appreciable for 5-year survival predicted to be very low or high.

21.6 Presenting the Model

To present point and interval estimates of predictor effects we draw a hazard ratio chart (Figure 21.4), and to make a final presentation of the model we draw a nomogram having multiple "predicted value" axes. Since the `ap` relationship is so non-monotonic, use a 20 : 1 hazard ratio for this variable.

```
plot(summary(f, ap=c(1,20)), log=TRUE, main='')  # Figure 21.4
```

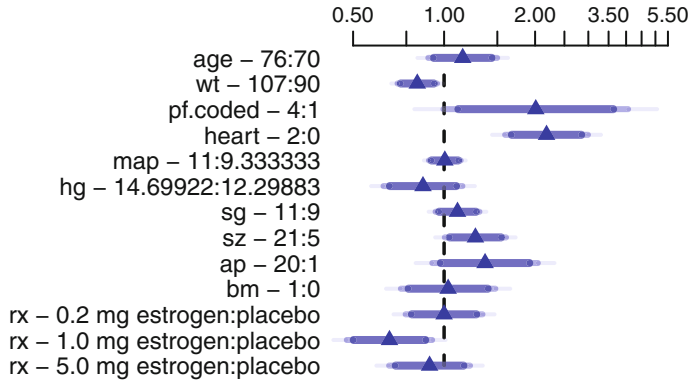

Fig. 21.4 Hazard ratios and multi-level confidence bars for effects of predictors in model, using default ranges except for `ap`

The ultimate graphical display for this model will be a nomogram relating the predictors to $X\hat{\beta}$, estimated three– and five-year survival probabilities and median survival time. It is easy to add as many "output" axes as desired to a nomogram.

```
surv  ← Survival(f)
surv3 ← function(x) surv(3*12,lp=x)
surv5 ← function(x) surv(5*12,lp=x)
quan  ← Quantile(f)
med   ← function(x) quan(lp=x)/12
ss    ← c(.05,.1,.2,.3,.4,.5,.6,.7,.8,.9,.95)

nom ← nomogram(f, ap=c(.1,.5,1,2,3,4,5,10,20,30,40),
                fun=list(surv3, surv5, med),
                funlabel=c('3-year Survival','5-year Survival',
                  'Median Survival Time (years)'),
                fun.at=list(ss, ss, c(.5,1:6)))
plot(nom, xfrac=.65, lmgp=.35)   # Figure 21.5
```

21.7 Problems

Perform Cox regression analyses of survival time using the Mayo Clinic PBC dataset described in Section 8.9. Provide model descriptions, parameter estimates, and conclusions.

1. Assess the nature of the association of several predictors of your choice. For polytomous predictors, perform a log-rank-type score test (or k-sample ANOVA extension if there are more than two levels). For continuous predictors, plot a smooth curve that estimates the relationship between the predictor and the log hazard or log–log survival. Use both parametric and nonparametric (using martingale residuals) approaches. Make a test of H_0 : predictor is not associated with outcome versus H_a : predictor

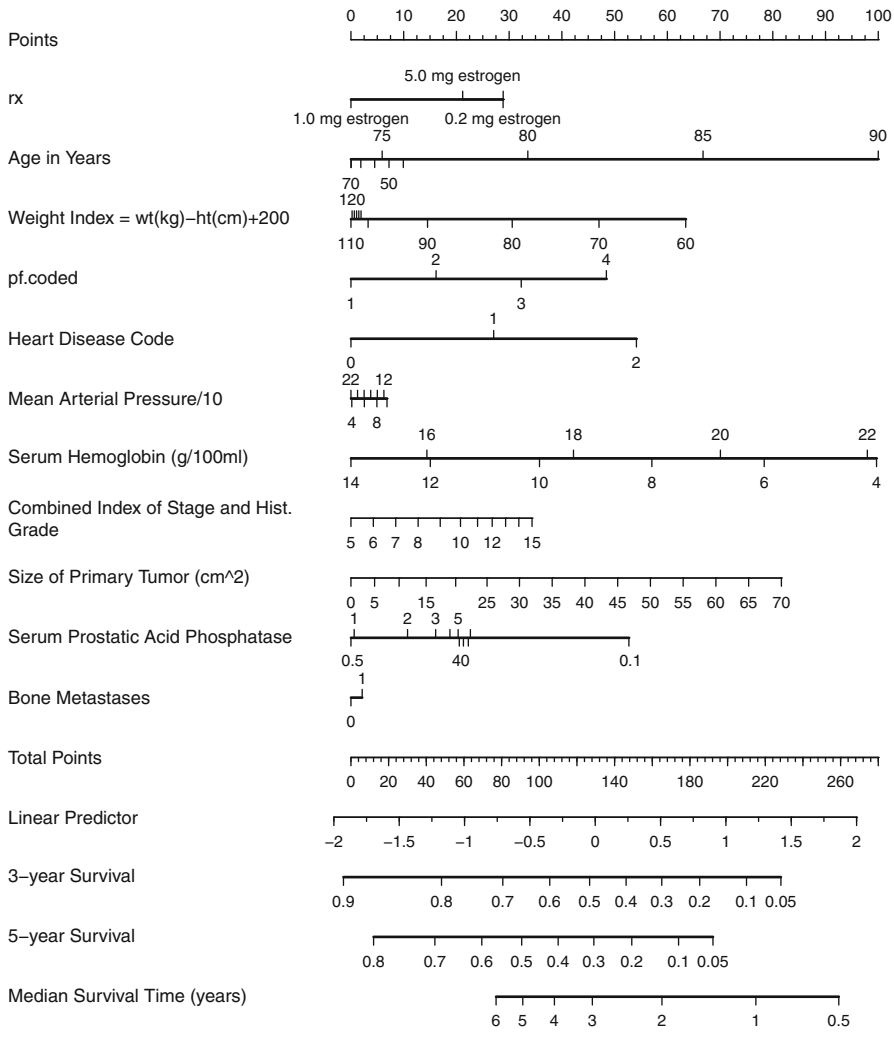

Fig. 21.5 Nomogram for predicting death in prostate cancer trial

is associated (by a smooth function). The test should have more than 1 d.f. If there is no evidence that the predictor is associated with outcome. Make a formal test of linearity of each remaining continuous predictor. Use restricted cubic spline functions with four knots. If you feel that you can't narrow down the number of candidate predictors without examining the outcomes, and the number is too great to be able to derive a reliable model, use a data reduction technique and combine many of the variables into a summary index.

2. For factors that remain, assess the PH assumption using at least two methods, after ensuring that continuous predictors are transformed to be as linear as possible. In addition, for polytomous predictors, derive log cumulative hazard estimates adjusted for continuous predictors that do not assume anything about the relationship between the polytomous factor and survival.
3. Derive a final Cox PH model. Stratify on polytomous factors that do not satisfy the PH assumption. Decide whether to categorize and stratify on continuous factors that may strongly violate PH. Remember that in this case you can still model the continuous factor to account for any residual regression after adjusting for strata intervals. Include an interaction between two predictors of your choosing. Interpret the parameters in the final model. Also interpret the final model by providing some predicted survival curves in which an important continuous predictor is on the x-axis, predicted survival is on the y-axis, separate curves are drawn for levels of another factor, and any other factors in the model are adjusted to specified constants or to the grand mean. The estimated survival probabilities should be computed at $t = 730$ days.
4. Verify, in an unbiased fashion, your "final" model, for either calibration or discrimination. Validate intermediate steps, not just the final parameter estimates.