10. Check distributional assumptions and choose a different model if needed (in the case of Cox models, stratification or time-dependent covariables can be used if proportional hazards is violated).
11. Do limited backwards step-down variable selection.[63] Note that since stepwise techniques do not really address overfitting and they can result in a loss of information, full model fits (that is, leaving all hypothesized variables in the model regardless of *P*-values) are frequently more discriminating than fits after screening predictors for significance.[2,40] They also provide confidence intervals with the proper coverage, unlike models that are reduced using a stepwise procedure,[60,64,65] from which confidence intervals are falsely narrow. A compromise would be to test a *pre-specified* subset of predictors, deleting them if their total $\chi^2 < 2 \times$ d.f. If the χ^2 is that small, the subset would likely not improve model accuracy.
12. This is the 'final' model.
13. Validate this model for calibration and discrimination ability, preferably using bootstrapping. Steps 7 to 11 must be repeated for each bootstrap sample, at least approximately. For example, if age was transformed when building the final model, and the transformation was suggested by the data using a fit involving age and age^2, each bootstrap repetition should include both age variables with a possible step-down from the quadratic to the linear model based on automatic significance testing at each step.
14. If doing stepwise variable selection, present a summary table depicting the variability of the list of 'important factors' selected over the bootstrap samples or cross-validations. This is an excellent tool for understanding why data-driven variable selection is inherently ambiguous.
15. Estimate the likely shrinkage of predictions from the model, either using equation (2) or by bootstrapping an overall slope correction for the predictions.[34] Consider shrinking the predictions to make them calibrate better, unless shrinkage was built–in. That way, a predicted 0·4 mortality is more likely to validate in a new patient series, instead of finding that the actual mortality is only 0·2 because of regression to the mean mortality of 0·1.

8. SOFTWARE

Modern statistical software such as S-Plus[37] on UNIX workstations makes it quite feasible to perform the extensive calculations required to do the recommended model building steps. The first author has written a package of UNIX S-Plus functions called Design[66] that allow the analyst to perform all analyses mentioned here including tests of linearity, pooled interaction tests, model validation and graphical methods for interpreting models. Here are some examples:

```
# First find optimum transformations relating each predictor to each
# other, and use multiple regression in these transformations to
# impute missing values. Use shrinkage to avoid over-imputing
trans ← transcan( ~ age + cholesterol + sys.bp + weight, imputed = T, shrink = T)
cholesterol ← impute(trans, cholesterol) # impute missings
sys.bp ← impute(trans, sys.bp)
# Fit a Cox P.H. model allowing some interactions with age and
# nonlinearity in cholesterol and sys.bp using restricted cubic splines
# x = T, y = T means store data in fit for future bootstrapping
fit ← cph(Surv(fu.time, death) ~ age * (rcs(cholesterol) + rcs(sys.bp)) +
                    weight, x = T, y = T, surv = T, time.inc = 5)
anova(fit)                          # automatic pooled Wald tests
fastbw(fit)                         # fast backward step-down
```

Table II. Candidate predictors and d.f.

Predictor	Name	Number of parameters	Original levels
Dose of oestrogen	rx	3	placebo, 0·2, 1·0, 5·0 mg oestrogen
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 disturbance, block, strain, old myocardial infarct, new MI
Serum haemoglobin (g/100 ml)	hg	3	
Tumour size (cm^2)	sz	3	
Stage/histologic grade combination	sg	3	
Serum prostatic acid phosphatase	ap	3	
Bone metastasis	bm	1	present/absent

```
# Next validate model, penalizing for backward stepdown variable selection
validate(fit, B = 100, bw = T)          # bootstrap validation of accuracy indexes
calibrate(fit, B = 100, bw = T, u = 5) # bias-corrected 5-yr survival calibration
plot(summary(fit))                      # plot hazard ratios with confidence limits
nomogram(fit)                           # draw nomogram displaying how model works
latex(fit)                              # typeset model equation
```

The Design library includes a function rcorr.cens for computing the general c–index, and the function val.prob which produced Figure 1 and also prints a variety of accuracy measures. For binary and ordinal logistic models and for ordinary linear models, Design has a general penalized maximum likelihood estimation facility. Design is available in the statlib repository (Internet address lib.stat.cmu.edu). transcan and impute are separate functions in statlib which work on UNIX as well as DOS Windows S-Plus. Some other software systems which have some intermediate-level capabilities include Stata (Computer Resources Center Inc., College Station TX), SPIDA (NHMRC Clinical Trials Centre, Eastwood, NSW Australia), and SAS (SAS Institute Inc., Cary NC).

9. CASE STUDY

Consider the 506-patient prostate cancer dataset from Byar and Green[67] which has also been analysed in references 68 and 69. The data are listed in reference 70, Table 46, and are available by Internet at utstat.toronto.edu in the directory /pub/data-collect. These data were from a randomized trial comparing four treatments for stage 3 and 4 prostate cancer, with almost equal numbers of patients on placebo and each of three doses of oestrogen. Four patients had missing values on all of the following variables: wt, pf, hx, sbp, dbp, ekg, hg, bm; two of these patients were also missing sz (see Table II for abbreviations). These patients will be excluded from consideration.

There are 354 deaths among the 502 patients. If we only wanted to test for a drug effect on survival time, a simple rank-based analysis would suffice. To be able to test for differential treatment effect or to estimate prognosis or expected absolute treatment benefit for individual

patients, however, we need a multivariable survival model.[3] First we consider fitting a full additive model which does not assume linearity of effect for any predictor. Categorical predictors will be expanded using dummy variables. For pf we could lump the last two categories since the last category has only two patients. Likewise, we could combine the last two levels of ekg. Continuous predictors will be expanded by fitting 4-knot restricted cubic spline functions, which contain two non-linear terms and thus have a total of 3 d.f. Table II 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. which should not be artificially reduced by 'univariable screening' or graphical assessments of association with death. This is about $\frac{1}{10}$ 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 (2). We use a Cox proportional hazards model for time until death. The UNIX S-Plus Design library fits the full model using restricted cubic spline expansions and makes use of Therneau's survival4 package in statlib[71] to perform the calculations. First we invoke the transcan function and impute functions (from statlib for any versions of S-Plus) to develop customized non-linear imputation equations for all predictors and to apply these equations to impute missing values.

```
# Define function for easy determination of whether a value is in a list
'%in%' ← function (a, b) match (a, b, nomatch = 0) > 0

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

pf.coded  ← as.integer(pf)   # save original pf, re-code to 1-4
levels(pf) ← c(levels(pf) [1 : 3], levels(pf) [3])  # combine last 2 levels of original
w  ← transcan( ~ sz + sg + ap + sbp + dbp + age + wt + hg +
                 ekg + pf + bm + hx, imputed = T, impcat = 'tree')
sz  ← impute(w, sz)        # uses imputation rule w
sg  ← impute(w, sg)
age ← impute(w, age)
wt  ← impute(w, wt)
ekg ← impute(w, ekg)

dd  ← datadist(rx, age, wt, pf, pf.coded, heart, map, hg, sz, sg, ap, bm)
options(datadist = 'dd')      # datadist stores characteristics of raw data

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(ap,4) + bm)
```

The likelihood ratio χ^2 statistic is 140 with 36 d.f. This test is highly significant so some modelling is warranted. The AIC value (on the χ^2 scale) is $140 - 2 \times 36 = 68$. The rough shrinkage estimate is 0·743 (104/140) so we estimate that 26% of the model fitting will be noise, especially with regard to calibration accuracy. The approach of reference 2 is to fit this full model and to shrink predicted values. We will 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 perform a variable clustering analysis which for our purposes we will do informally. The data reduction strategy is listed in Table III. For ap, more exploration is desired to be able to model the shape of effect with such a highly skewed

Table III. Data reduction strategy (blinded to Y)

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 and benign, 1 = either, 0 = none	5
sbp, dbp	Combine into mean arterial bp and use 3 knots: map = $\frac{2}{3}$ dpb + $\frac{1}{3}$ spb	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 in spline to achieve numerical stability: add 2 knots	−2

distribution. Since we expect the tumour variables to be strong prognostic factors we will retain them as separate variables. No assumption will be made for the dose-response shape for oestrogen, as there was reason to expect a non-monotonic effect due to competing risks for cardiovascular death.

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

f ← cph(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),6) + bm,
        x = T, y = T, surv = T, time.inc = 5*12)
# x, y for predict, validate, calibrate; surv, time.inc for calibrate
```

The total savings is thus 11 d.f. The likelihood ratio χ^2 is 126 with 25 d.f., with a slightly improved AIC of 76. The rough shrinkage estimate is slightly better at 0·80, but still worrisome. A further data reduction might be achieved by using the transcan transformations determined from self-consistency of predictors, but we will stop here and use this model.

Now assess this model in more detail by examining coefficients and summarizing multiple parameters within predictors using Wald statistics.

```
f      # writing an object name in S causes it to be printed

Cox Proportional Hazards Model

cph(formula = S ~ rx + rcs(age, 4) + rcs(wt, 3) + pf.coded + heart + rcs(map, 3) +
    rcs(hg, 4) + rcs(sz, 3) + rcs(sg, 3) + rcs(log(ap), 6) + bm,
    x = T, y = T, surv = T, time.inc = 5 * 12)

Obs Events Model L.R. d.f. P Score Score P R2
502    354    126  25  0  135    0  0·221

                          coef se(coef)          z          p
rx = 0·2 mg estrogen   3.74e − 03 1.50e − 01  0·0250  9·80e − 01
rx = 1.0 mg estrogen  −4·21e − 01 1·66e − 01 −2·5427  1·10e − 02
rx = 5.0 mg estrogen  −9·73e − 02 1·58e − 01 −0·6176  5·37e − 01
                age   −1·17e − 02 2·35e − 02 −0·4995  6·17e − 01
               age'    2·00e − 02 3·86e − 02  0·5190  6·04e − 01
```

age''	2·71e − 01	4·95e − 01	0·5482	5·84e − 01
wt	−2·46e − 02	9·39e − 03	−2·6175	8·86e − 03
wt'	1·84e − 02	1·12e − 02	1·6379	1·01e − 01
pf.coded	2·25e − 01	1·21e − 01	1·8625	6·25e − 02
heart	4·18e − 01	8·08e − 02	5·1723	2·31e − 07
map	3·24e − 02	8·49e − 02	0·3817	7·03e − 01
map'	−4·57e − 02	9·41e − 02	−0·4857	6·27e − 01
hg	−1·56e − 01	7·68e − 02	−2·0343	4·19e − 02
hg'	7·42e − 02	2·10e − 01	0·3530	7·24e − 01
hg''	5·08e − 01	1·27e + 00	0·4014	6·88e − 01
sz	1·00e − 02	1·44e − 02	0·6955	4·87e − 01
sz'	8·79e − 03	2·37e − 02	0·3715	7·10e − 01
sg	7·19e − 02	7·86e − 02	0·9138	3·61e − 01
sg'	−7·04e − 03	9·83e − 02	−0·0716	9·43e − 01
ap	−7·96e − 01	3·11e − 01	−2·5584	1·05e − 02
ap'	4·89e + 01	2·18e + 01	2·2482	2·46e − 02
ap''	−3.64e + 02	1·59e + 02	−2·2909	2·20e − 02
ap'''	4·04e + 02	1·75e + 02	2·3057	2·11e − 02
ap''''	−9·69e + 01	4·16e + 01	−2·3311	1·97e − 02
bm	3·25e − 02	1·81e − 01	0·1790	8·58e − 01

```
# The terms with ', '', etc. after the name are cubic spline nonlinear terms
# The dose effect is apparently nonlinear.

anova(f)      # output was actually typesetted automatically using latex(anova(f))
              # latex requires the print.display package from statlib
```

There are 12 parameters associated with non-linear effects, and the overall test of linearity indicates the strong presence of non-linearity for at least one of the variables age, wt, map, hg, sz, sg, ap (see Table IV). There is a difference in survival time between at least two of the doses of oestrogen.

Now that we have a tentative model, let us examine the model's distributional assumptions. As mentioned in Section 4.3, the Schoenfeld partial residuals are an effective tool for checking the proportional hazards assumption in the Cox model. Grambsch and Therneau[72] have modified these residuals so that smoothed plots of them estimate the effect of predictors on the log instantaneous hazard rate as a function of follow-up time. Their scaled residuals estimate $\beta(t)$, the regression coefficient as a function of time. 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 the terms in the model to transform that predictor to be optimally related 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 oestrogen, for example, we code the effect as 0 (placebo), 0·0037 (0·2 mg), −0·421 (1.0 mg), and −0·0973 (5.0 mg), and age is transformed as −0·0117 age + 0·02 age' + 0·271 age'', which in most simple form is

$$-1{\cdot}17\times10^{-2}\text{age} + 3{\cdot}48\times10^{-5}(\text{age}-56)_+^3 + 4{\cdot}71\times10^{-4}(\text{age}-71)_+^3$$

$$-1{\cdot}01\times10^{-3}(\text{age}-75)_+^3 + 5{\cdot}09\times10^{-4}(\text{age}-80)_+^3$$

where $(x)_+$ means to ignore that term if $x \leqslant 0$, and the knots for age are 56, 71, 75 and 80 years.

In S-Plus 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

Table IV. Wald statistics for S

	χ^2	d.f.	P
rx	8·38	3	0·0387
age	12·85	3	0·0050
Non-linear	8·18	2	0·0168
wt	8·87	2	0·0118
Non-linear	2·68	1	0·1014
pf.coded	3·47	1	0·0625
heart	26·75	1	<0·0001
map	0·25	2	0·8803
Non-linear	0·24	1	0·6272
hg	11·85	3	0·0079
Non-linear	6·92	2	0·0314
sz	10·60	2	0·0050
Non-linear	0·14	1	0·7102
sg	3·14	2	0·2082
Non-linear	0·01	1	0·9429
ap	13·17	5	0·0218
Non-linear	12·93	4	0·0116
bm	0·03	1	0·8579
TOTAL NON-LINEAR	**30·28**	**12**	**0·0025**
TOTAL	**128·08**	**25**	**<0·0001**

formulae.

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

The fit f.short based on the matrix z of single d.f. predictors has the same LR χ^2 of 126 as the fit f, but with a falsely low 11 d.f. All regression coefficients are unity.

Now get scaled Schoenfeld residuals separately for each predictor and test the proportional hazards assumption for each using the 'correlation with time' test. Also plot smoothed trends in the residuals. The plot method for cox.zph objects uses restricted cubic splines to smooth the relationship.

```
phtest ← cox.zph(f.short, transform = 'identity')
phtest
             rho      chisq       p
rx       0·12965     6·5451  0·0105
age     −0·08911     2.8518  0·0913
wt      −0·00878     0·0269  0·8697
pf.coded −0·06238    1·4278  0·2321
heart    0·01017     0·0451  0·8319
map      0·03928     0·4998  0·4796
hg      −0·06678     1·7368  0·1876
sz      −0·05262     0·9834  0·3214
sg      −0·04276     0·6474  0·4210
ap       0·01237     0·0558  0·8133
bm       0·04891     0·9241  0·3364
GLOBAL      NA      15·3776  0·1659
```

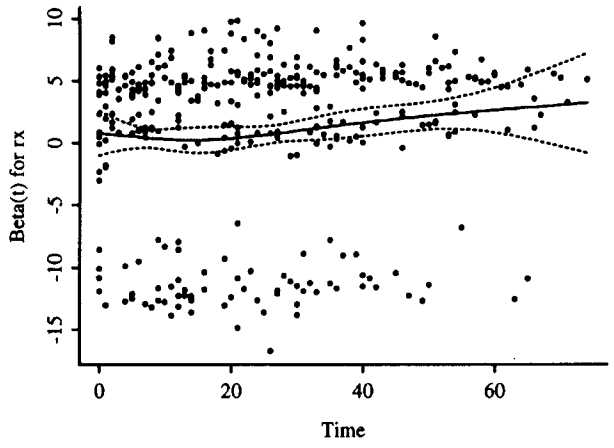

Figure 2. Raw and spline-smoothed scaled Schoenfeld residuals for dose of oestrogen, non-linearly coded from the Cox model fit, with ± 2 standard errors.[71]

Only the drug effect significantly changes over time ($P = 0{\cdot}01$ 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 0·17. A graphical examination of the trends does not find anything interesting for the last 10 variables. A residual plot is drawn for rx alone and is shown in Figure 2.

```
plot(phtest, var = 'rx')
```

We will ignore the possible increase in effect of oestrogen 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.

Note that the model has several insignificant predictors. These will not be deleted, as that would not improve predictive accuracy and it would make confidence intervals for $\hat{\beta}$ or for predicted survival probabilities with the correct coverage probabilities hard to obtain.[64] At this point it would be reasonable to test pre-specified interactions. Here we will 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 will do approximate tests on the data-driven codings 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)
anova(f.ia)
```

Factor	Chi-Square	d.f.	P
z.dose (Factor + Higher Order Factors)	18·9	11	0·062
All Interactions	12·2	10	0·273
z.other (Factor + Higher Order Factors)	134·3	20	0·000
All Interactions	12·2	10	0·273
z.dose * z.other (Factor + Higher Order Factors)	12·2	10	0·273
TOTAL	137·3	21	0·000