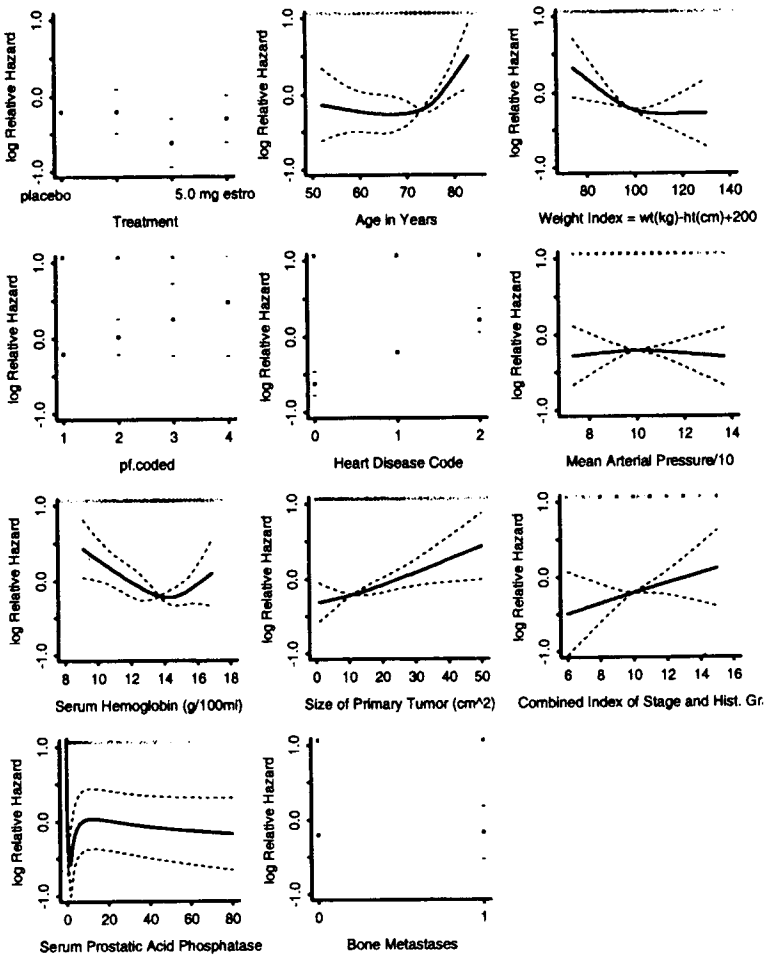

Figure 3. Shape of each predictor on log hazard of death. Y-axis shows $X\hat{\beta}$, but the predictors not plotted are set to reference values. 'Rug plots' on the top of each graph show the data density of the predictor. Note the highly non-monotonic relationship with ap, and the increased slope after age 70 which has been found in outcome models for various diseases

Here 'Factor + Higher Order Factors' means the combined main effect and interaction effect. The global test of additivity has $P = 0{\cdot}27$, so we will ignore the interactions (and also forget to penalize for having looked for them below!).

The following UNIX S-Plus statements plot how each predictor is related to the log hazard of death, along with 0·95 confidence bands. Note 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).

```
par(mfrow = c(3, 4))      # 4 x 3 matrix of graphs
r ← c(-1, 1)              # use common y-axis range for all
plot(f, rx = NA,          ylim = r)   NA → use default range for predictor
plot(f, age = NA,         ylim = r)
scatld(age)                           # scatld from statlib, for any S-Plus
plot(f, wt = NA,          ylim = r)   # scatld shows data density
...
```

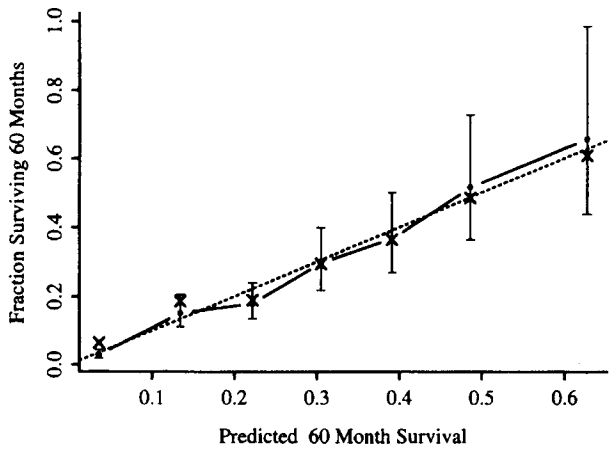

Figure 4. Bootstrap estimate of calibration accuracy for 5-year estimates from the final Cox model. Dots correspond to apparent predictive accuracy. × marks the bootstrap-corrected estimates

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 200 re-samples) to penalize for possible overfitting, as discussed in Section 6.

```
validate(f, B = 200, dxy = T, pr = T)
```

	index.orig	training	test	optimism	index. corrected	n
Dxy	−0·337377	−0·364644	−0·30976	−0·05488	−0·28250	200
R2	0·221444	0·261369	0·18445	0·07691	0·14453	200
Slope	1.000000	1.000000	0·78464	0·21536	0·78464	200

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 −0·34, but a better estimate of how well the model will discriminate prognoses in the future is $D_{xy} = -0{\cdot}28$. The bootstrap estimate of slope shrinkage is 0·78, surprisingly 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 5 years. As detailed in Section 5, the bootstrap is used to estimate the optimism in how well predicted 5-year survival from the final Cox model tracks Kaplan–Meier 5-year estimates, stratifying by grouping patients in subsets with about 70 patients per interval of predicted 5-year survival.

```
plot(calibrate(f, B = 200, u = 5 * 12, m = 70))
```

The estimated calibration curves are shown in Figure 4. Bias-corrected calibration is very good except for the two groups with extremely bad prognosis – their survival is slightly better than predicted, consistent with regression to the mean. Even there, the absolute error is low despite a large relative error. Hence for this example it may not be worthwhile to develop a model using shrinkage.

Now compare this analysis with three previous analyses of this dataset. In all three analyses, all continuous covariables were arbitrarily categorized into intervals and scored with somewhat arbitrary category codes. In none of the three were `sbp`, `dbp`, `ekg`, `ap`, `bm` considered. Patients having missing values on any of the candidate predictors were excluded from consideration.

Turn first to Byar and Green,[67] who used an exponential survival model and dichotomized treatment by combining placebo and low dose and combining the two highest doses. The important predictors were found to be hx, sg, sz, hg, and the following interactions were detected in an exploratory analysis which did not control for multiple comparisons: rx × sg and rx × age. These interactions were not significant in the present model (even if dose were re-coded as in Byar and Green).

Kay[68] considered Cox models for various causes of death. For time until all-cause mortality, Kay found that the most important predictors were sz, hx, sg, age. The treatment along with age, hx were significant predictors of cardiovascular death. The treatment (in the opposite direction), and hg, sz, sg predicted cancer death. Treatment and age, wt predicted time until death from other causes.

Sauerbrei and Schumacher[69] also used a Cox model and an approach in which a backward elimination procedure was done for each of 100 bootstrap samples. The relative frequency of selection of variables as 'important' was used as the criterion for inclusion of variables in the final model. Variables were retained if they were selected $\geqslant 70$ times. All candidate predictors met this criterion. Treatment interactions involving age and sg were the most common interactions (56 and 48 bootstrap repetitions, respectively), but they did not meet the criterion for selection. The authors noted that these interactions were misleadingly more significant in a model which only adjusted for 'significant' predictors instead of all candidate predictors.

None of the three references just cited provided a model validation or quantified the predictive discrimination of the final model.

10. SUMMARY

Methods were described for developing clinical multivariable prognostic models and for assessing their calibration and discrimination. A detailed examination of model assumptions and an unbiased assessment of predictive accuracy will uncover problems that may make clinical prediction models misleading or invalid. The modelling strategy presented in Section 7 provides one sequence of steps for avoiding the pitfalls of multivariable modelling so that its many advantages can be realized.

ACKNOWLEDGEMENTS

This work was supported by research grants HL-17670, HL-29436, HL-36587, HL-45702 and HL-09315 from the National Heart, Lung and Blood Institute, Bethesda, Maryland, research grants HS-03834, HS-05635, HS-06503, HS-06830, and HS-07137 from the Agency for Health Care Policy and Research, Rockville, Maryland, and grants from the Robert Wood Johnson Foundation, Princeton, NJ.

REFERENCES

1. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A., 'Evaluating the yield of medical tests', *Journal of the American Medical Association*, **247**, 2543–2546 (1982).
2. Spiegelhalter, D. J. 'Probabilistic prediction in patient management', *Statistics in Medicine*, **5**, 421–433 (1986).
3. Knaus, W. A., Harrell, F. E., Fisher, C. J., Wagner, D. P., Opan, S. M., Sadoff, J. C., Draper, E. A., Walawander, C. A., Conboy, K. and Grasela, T. H. 'The clinical evaluation of new drugs for sepsis: A prospective study design based on survival analysis', *Journal of the American Medical Association*, **270**, 1233–1241 (1993).

4. Donner, A. 'The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values', *American Statistician*, **36**, 378–381 (1982).
5. Roberts, J. S. and Capalbo, G. M. 'A SAS macro for estimating missing values in multivariate data', *Proceedings of the Twelfth Annual SAS Users Group International Conference*, (Cary NC), SAS Institute, 939–941 (1987).
6. Buck, S. F. 'A method of estimation of missing values in multivariate data suitable for use with an electronic computer', *Journal of the Royal Statistical Society, Series B*, **22**, 302–307 (1960).
7. Timm, N. H. 'The estimation of variance-covariance and correlation matrices from incomplete data', *Psychometrika*, **35**, 417–437 (1970).
8. Kuhfeld, W. F. 'The PRINQUAL procedure', in *SAS/STAT User's Guide*, 4th edn., vol. **2**, SAS Institute, Cary NC, 1990, chapter 34, pp. 1265–1323.
9. Schemper, M. 'Non-parametric analysis of treatment-covariate interaction in the presence of censoring', *Statistics in Medicine*, **7**, 1257–1266 (1988).
10. Cox, D. R. 'The regression analysis of binary sequences (with discussion)', *Journal of the Royal Statistical Society, Series B*, **20**, 215–242 (1958).
11. Walker, S. H. and Duncan, D. B. 'Estimation of the probability of an event as a function of several independent variables', *Biometrika*, **54**, 167–178 (1967).
12. van Houwelingen J. C. and le Cessie, S. 'Logistic regression, a review', *Statistica Neerlandica*, **42**, 215–232 (1988).
13. Collett, D. *Modelling Binary Data*. Chapman and Hall, London, 1991.
14. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
15. Collett, D. *Modelling Survival Data in Medical Research*. Chapman and Hall, London 1994.
16. Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Wiley, New York 1982.
17. Derksen S. and Keselman, H. J. 'Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables', *British Journal of Mathematical and Statistical Psychology*, **45**, 265–282 (1992).
18. Grambsch, P. M. and O'Brien, P. C. 'The effects of transformations and preliminary tests for non-linearity in regression', *Statistics in Medicine*, **10**, 697–709 (1991).
19. Verweij, P. and van Houwelingen, H. C. 'Penalized likelihood in Cox regression', *Statistics in Medicine*, **13**, 2427–2436 (1994).
20. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. 'Regression modelling strategies for improved prognostic prediction', *Statistics in Medicine*, **3**, 143–152 (1984).
21. Marshall, G., Grover, F. L., Henderson, W. G. and Hammermeister, K. E. 'Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery', *Statistics in Medicine*, **13**, 1501–1511 (1994).
22. Jolliffe, I. T. *Principal Component Analysis*, Springer-Verlag, New York, 1986.
23. Jackson, J. E. *A User's Guide to Principal Components*, Wiley, New York, 1991.
24. Smith, L. R., Harrell, F. E. and Muhlbaier, L. H. 'Problems and potentials in modelling survival', in: Grady, M. L. and Schwartz, H. A. (eds.), *Medical Effectiveness Research Data Methods (Summary Report)*, *AHCPR Pub. No. 92–0056* US Dept. of Health and Human Services, Agency for Health Care Policy and Research, Rockville, Maryland, 1992, pp. 151–159.
25. Durrleman, S. and Simon, R. 'Flexible regression models with cubic splines', *Statistics in Medicine*, **8**, 551–561 (1989).
26. Harrell, F. E., Lee, K. L. and Pollock, B. G. 'Regression models in clinical studies: determining relationships between predictors and response', *Journal of the National Cancer Institute*, **80**, 1198–1202 (1988).
27. Sleeper, L. A. and Harrington, D. P. 'Regression splines in the Cox model with application to covariate effects in liver disease', *Journal of the American Statistical Association*, **85**, 941–949 (1990).
28. Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. 'Graphical methods for assessing logistic regression models (with discussion)', *Journal of the American Statistical Association*, **79**, 61–83 (1984).
29. Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*. Wiley, New York, 1989.
30. Therneau, T. M., Grambsch, P. M. and Fleming, T. R. 'Martingale-based residuals for survival models', *Biometrika*, **77**, 216–218 (1990).
31. Schoenfeld, D. 'Partial residuals for the proportional hazards regression model', *Biometrika*, **69**, 239–241 (1982).

32. Pettitt A. N. and Bin Daud, I. 'Investigating time dependence in Cox's proportional hazards model', *Applied Statistics*, **39**, 313–329 (1990).
33. Harrell, F. E., Pollock, B. G. and Lee, K. L. 'Graphical methods for the analysis of survival data', in *Proceedings of the Twelfth Annual SAS Users Group International Conference*, Cary, NC, pp. 1107–1115, SAS Institute, Inc., 1987.
34. van Houwelingen, J. C. and le Cessie, S. 'Predictive value of statistical models', *Statistics in Medicine*, **8**, 1303–1325 (1990).
35. Friedman, J. H. 'A variable span smoother', Technical Report 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984.
36. Cleveland, W. S. 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, **74**, 829–836 (1979).
37. Statistical Sciences, *S-Plus User's Manual, Version 3.2.*, StatSci, a division of MathSoft, Inc., Seattle WA, 1993.
38. Brier, G. W. 'Verification of forecasts expressed in terms of probability,' *Monthly Weather Review*, **75**, 1–3 (1950).
39. Kaplan, E. L. and Meier, P. 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, **53**, 457–481 (1958).
40. Copas, J. B. 'Regression, prediction and shrinkage (with discussion)', *Journal of the Royal Statistical Society, Series B*, **45**, 311–354 (1983).
41. Copas, J. B. 'Cross–validation shrinkage of regression predictors', *Journal of the Royal Statistical Society, Series B*, **49**, 175–183 (1987).
42. Gray, R. J. 'Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis', *Journal of the American Statistical Association*, **87**, 942–951 (1992).
43. Goodman, L. A. and Kruskal, W. H. *Measures of Association for Cross-Classifications*, Springer-Verlag, New York 1979.
44. Brown, B. W., Hollander, M. and Korwar, R. M. 'Nonparametric tests of independence for censored data, with applications to heart transplant studies', in: Proschan, F. and Serfling,R. J. (eds), *Reliability and Biometry*, SIAM, Philadelphia, 1974.
45. Schemper, M. 'Analyses of associations with censored data by generalized Mantel and Breslow tests and generalized Kendall correlation', *Biometrical Journal*, **26**, 309–318 (1984).
46. Bamber, D. 'The area above the ordinal dominance graph and the area below the receiver operating characteristic graph', *Journal of Mathematical Psychology*, **12**, 387–415 (1975).
47. Hanley, J. A. and McNeil, B. J. 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, **143**, 29–36 (1982).
48. Liu, K. and Dyer, A. R. 'A rank statistic for assessing the amount of variation explained by risk factors in epidemiologic studies', *American Journal of Epidemiology*, **109**, 597–606 (1979).
49. Nagelkerke, N. J. D. 'A note on a general definition of the coefficient of determination', *Biometrika*, **78**, 691–692 (1991).
50. Lee, K. L., Pryor, D. B., Harrell, F. E., Califf, R. M., Behar, V. S., Floyd, W. L., Morris, J. J., Waugh, R. A., Whalen, R. E. and Rosati, R. A. 'Predicting outcome in coronary disease: Statistical models versus expert clinicians', *American Journal of Medicine*, **80**, 553–560 (1986).
51. Harrell, F. E. and Lee, K. L. 'A comparison of the *discrimination* of discriminant analysis and logistic regression under multivariate normality', in Sen, P. K. (ed), *Biostatistics: Statistics in Biomedical, Public Health, and Environmental Sciences. The Bernard G. Greenberg Volume*, North-Holland, New York, 1985, pp. 333–343.
52. Korn, E. L. and Simon, R. 'Measures of explained variation for survival data', *Statistics in Medicine*, **9**, 487–503 (1990).
53. Picard, R. R. and Berk, K. N. 'Data splitting', *American Statistician*, **44**, 140–147 (1990).
54. Efron, B. 'Estimating the error rate of a prediction rule: improvement on cross-validation', *Journal of the American Statistical Association*, **78**, 316–331 (1983).
55. Linnet, K. 'Assessing diagnostic tests by a strictly proper scoring rule', *Statistics in Medicine*, **8**, 609–618 (1989).
56. Efron, B. and Gong, G. 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *American Statistician*, **37**, 36–48 (1983).
57. Efron, B. 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association*, **81**, 461–470 (1986).
58. Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

59. Roecker, E. B. 'Prediction error and its estimation for subset-selected models', *Technometrics*, **33**, 459–468 (1991).
60. Breiman, L. 'The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error', *Journal of the American Statistical Association*, **87**, 738–754 (1992).
61. Atkinson, A. C. 'A note on the generalized information criterion for choice of a model', *Biometrika*, **67**, 413–418 (1980).
62. Crawford, S. L., Tennstedt, S. L. and McKinlay, J. B. 'A comparison of analytic methods for non-random missingness of outcome data', *Journal of Clinical Epidemiology*, **48**, 209–219 (1995).
63. Mantel, N. 'Why stepdown procedures in variable selection', *Technometrics*, **12**, 621–625 (1970).
64. Altman, D. G. and Andersen, P. K. 'Bootstrap investigation of the stability of a Cox regression model', *Statistics in Medicine*, **8**, 771–783 (1989).
65. Hurvich, C. M. and Tsai, C. L. 'The impact of model selection on inference in linear regression', *American Statistician*, **44**, 214–217 (1990).
66. Harrell, F. E. 'Design: S-Plus functions for biostatistical/epidemiologic modelling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. Programs available from statlib@lib.stat.cmu.edu. Send E–mail 'send design from S", 1994.
67. Byar, D. P. and Green, S. B. 'The choice of treatment for cancer patients based on covariate information: application to prostate cancer', *Bulletin Cancer*, Paris, **67**, 477–488 (1980).
68. Kay, R. 'Treatment effects in competing-risks analysis of prostate cancer data', *Biometrics*, **42**, 203–211 (1986).
69. Sauerbrei, W. and Schumacher, M. 'A bootstrap resampling procedure for model building: Application to the Cox regression model', *Statistics in Medicine*, **11**, 2093–2109, (1992).
70. Andrews, D. F. and Herzberg, A. M. *Data*. New York, Springer-Verlag, 1985.
71. Therneau, T. 'Survival4: S functions for survival analysis. Programs available from statlib@lib.stat.cmu.edu. Send E–mail 'send survival4 from S,' 1995.
72. Grambsch, P. and Therneau, T. 'Proportional hazards tests and diagnostics based on weighted residuals', *Biometrika*, **81**, 515–526 (1994).