Regression Modeling and Validation Strategies

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
School of Medicine, University of Virginia

June, 1997

Department of Biomedical Informatics
Ljubljana University,
Ljubljana, Slovenia
Organizer: Janez Stare

Outline

- Finding Transformations for Continuous Predictors
- Aspects of a Model's Predictive Accuracy
- Perils of Overfitting
- Pitfalls of Stepwise Variable Selection
- Methods of Validating Models
- Graphical Depiction of Models

Finding Transformations

- Multivariable regression models assume that predictors relate linearly to some function of the responses
- No reason for nature to be so nice
- Can try different transformations, e.g., $\log$, $\sqrt{}$
- Can add nonlinear terms to model
- Example: fit a model containing age and square of age

  Allows parabolic (quadratic) shape for age effect
- More flexible: piecewise polynomials (spline functions)[17]

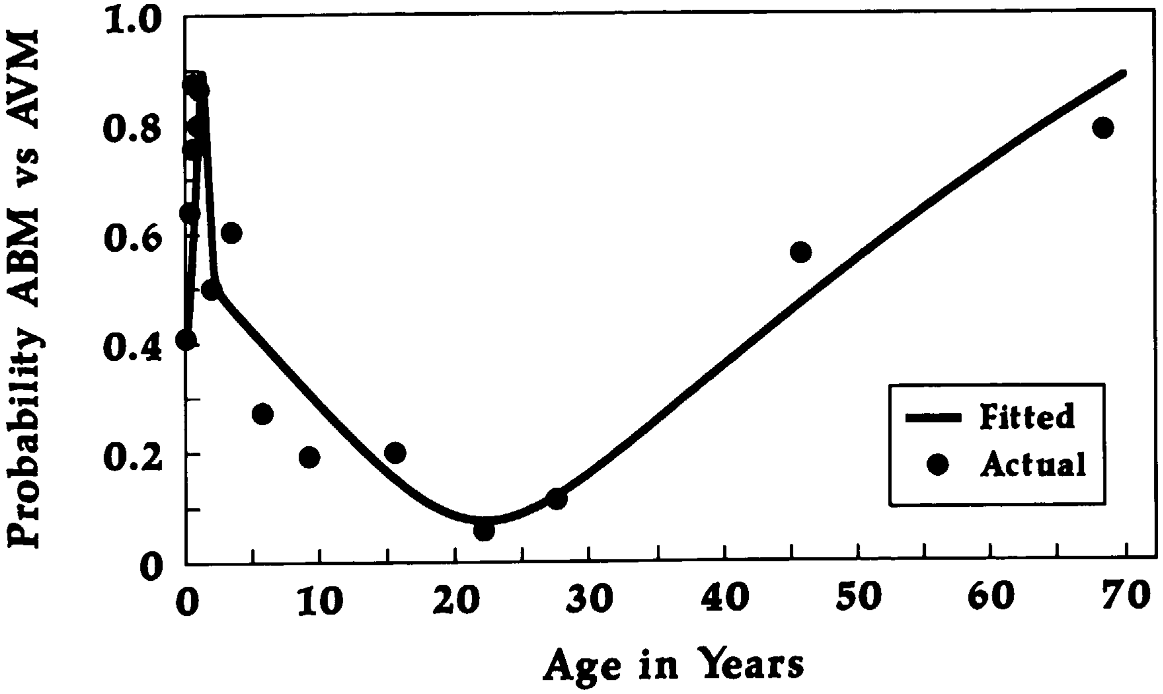


Figure 1: Linear spline fit for probability of bacterial vs. viral meningitis as a function of age at onset [29].

Predictive Accuracy

- Some models are used only for hypothesis testing
- If used for prediction, need to consider accuracy of predictions
- Calibration: observed responses agree with predicted responses
- Discrimination: model is able, through the use of predicted responses, to separate subjects with low observed responses from those with high responses

Perils of Overfitting

- Fitting a model with 20 patients and 20 variables (counting the intercept) will result in $R^2 = 1$ no matter what the variables are
- Analyzing too many variables for the available sample size will not cause a problem with *apparent* predictive accuracy
- Calibration or discrimination accuracy assessed on a new sample will suffer
- Caused by multiple comparison problems and trying to estimate too many parameters (regression coefficients) from the sample
- To use standard statistical methods, need to have a certain number of subjects per candidate predictor[a]

[a] The term *candidate* is used because one needs to count all variables examined for association with the response even if some of them are not included in the final model. This is because stepwise significance testing involves multiple comparison problems. See *regression to the mean* in the glossary handout.

variable for model to be able to validate on new data

- Continuous response: 10–20 subjects per candidate predictor
- Binary response: 10–20 subjects per less frequent of the two response values
- Survival analysis (time to event data): 10–20 subjects per event

Stepwise Variable Selection

- Add variables to a model according to statistical significance
- Commonly used, gives concise models
- Prone to problems of overstating importance of variables which are retained in the model
- Does not solve the “too many variables, too few subjects” problem, because “insignificant” variables are dropped on the basis of apparent lack of association
- Treating final model as if it were pre–specified can cause statistical problems (inflate type I error, confidence intervals too short)[15]

Validation Methods

- Need to use some validation method to honestly assess the likely performance of a model on a new series of subjects
- Data–splitting: split sample into two parts at random
  Use first part to develop model
  Use second part to measure predictive accuracy
- Is an honest method but assessment can vary greatly when take different splits
- Cross–validation: e.g., leave out $\frac{1}{10}$ of subjects, develop model in $\frac{9}{10}$, evaluate in $\frac{1}{10}$, repeat 10 times and average
- Still not very precise way to measure accuracy
- Bootstrap method is more precise and doesn't require holding back data (see glossary)

Validation Example

- Dataset of 200 observations on 20 *random* predictors
- Response variable is survival time, generated at random, independently of predictors
- Apparent calibration accuracy assessed by dividing observations into quintiles of predicted 0.5 year survival
- Fit 20 predictors to 100 events
- Fair *apparent* agreement between predicted and Kaplan–Meier survival over all strata (dots)
- Bias–corrected (overfitting–corrected) calibrations (Xs) gives accurate estimates (predicted survival probability actually unrelated to survival time)

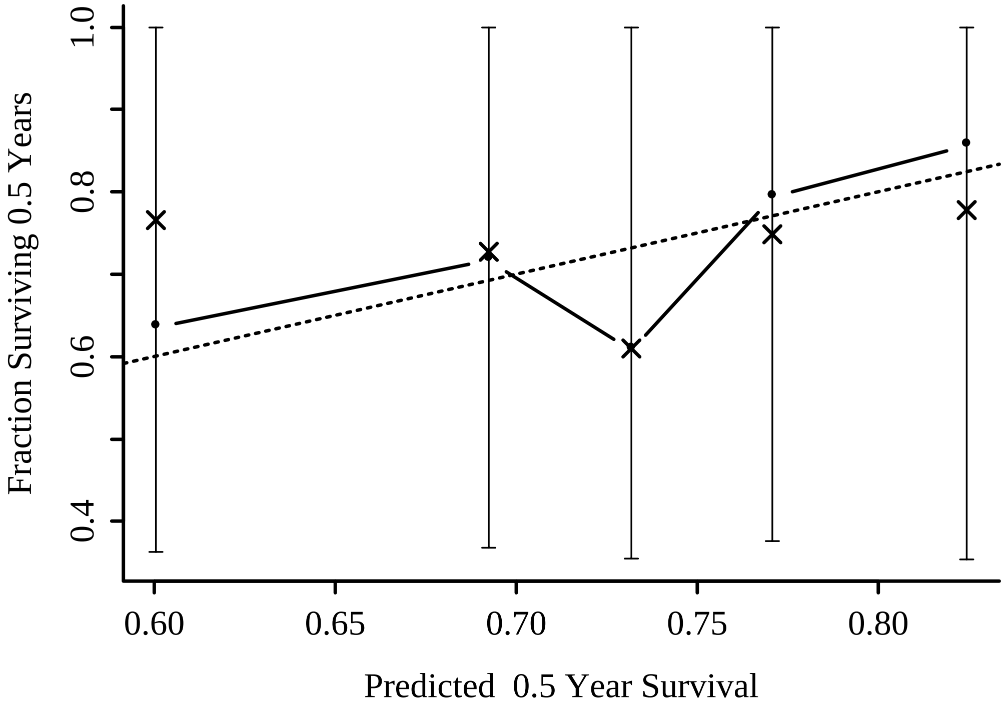
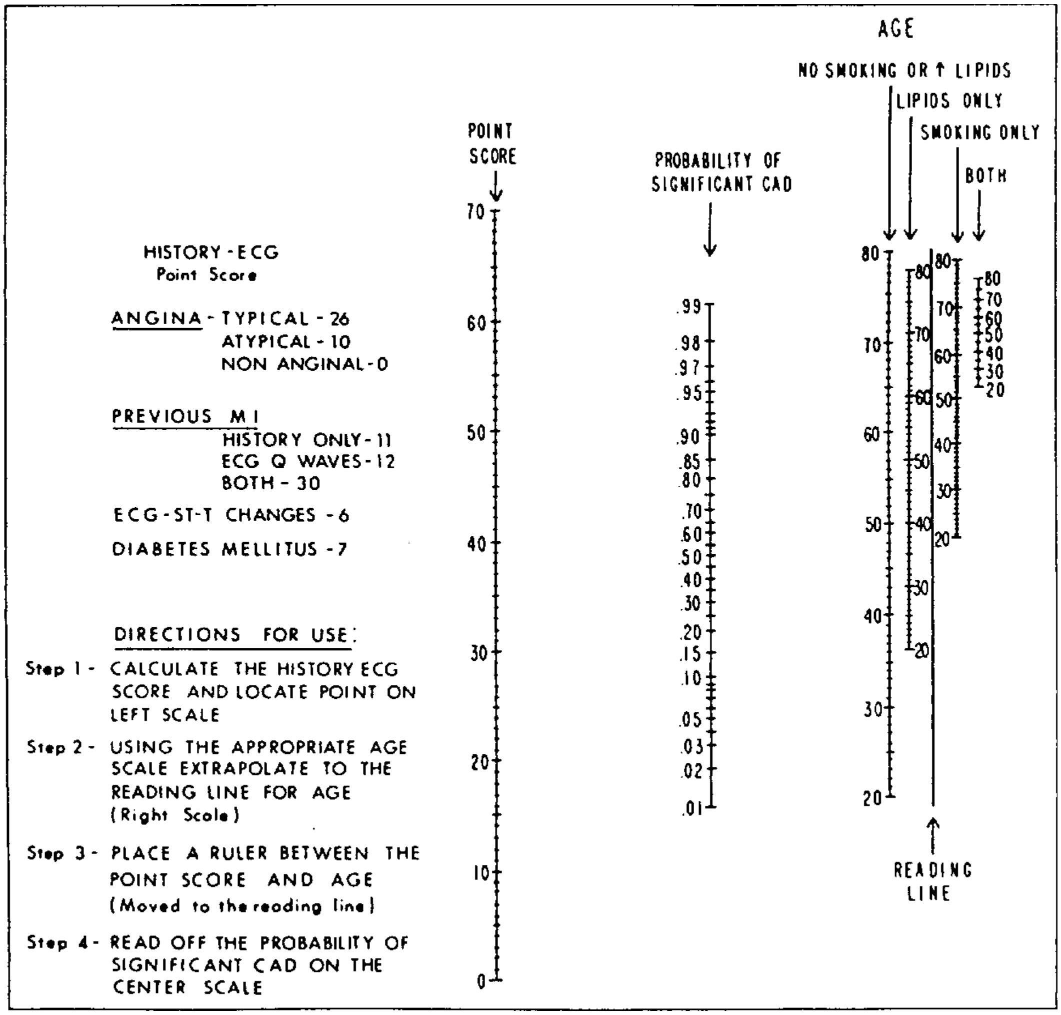


Figure 2: Calibration of random predictions using Efron's bootstrap with 50 re-samples and 40 patients per interval. Dataset has n=200, 100 uncensored observations, 20 random predictors, $\chi^2_{20} = 9.87$. ●: apparent calibration; X: bias–corrected calibration.

Graphical Depictions

- Model needn't be a black box
- Instead of concentrating on regression coefficients can draw effects of predictors each on its own axis in a nomogram
- Nonlinear effects will have nonlinear scales
- Each predictor put on a common scale but scales are labeled in the original scale of the predictor
- Addition of effects of predictors can be done by connecting two axes and seeing where the line hits a reading line, or by having a "points" axis and making the user manually add up the points each predictor receives



HISTORY-ECG
Point Score

ANGINA - TYPICAL - 26
ATYPICAL - 10
NON ANGINAL - 0

PREVIOUS MI
HISTORY ONLY - 11
ECG Q WAVES - 12
BOTH - 30

ECG-ST-T CHANGES - 6

DIABETES MELLITUS - 7

DIRECTIONS FOR USE:

Step 1 - CALCULATE THE HISTORY ECG SCORE AND LOCATE POINT ON LEFT SCALE

Step 2 - USING THE APPROPRIATE AGE SCALE EXTRAPOLATE TO THE READING LINE FOR AGE (Right Scale)

Step 3 - PLACE A RULER BETWEEN THE POINT SCORE AND AGE (Moved to the reading line)

Step 4 - READ OFF THE PROBABILITY OF SIGNIFICANT CAD ON THE CENTER SCALE

Figure 3: A nomogram for estimating the likelihood of significant coronary artery disease (CAD) in women. Depiction of a fitted binary logistic regression model. Categorical predictors have their points added manually. ECG = electrocardiographic; MI = myocardial infarction [25]. Presence of important age × risk factor interactions is handled by constructing separate age scales for each level of the interacting factor. Here, interaction means a change in the slope (regression coefficient) for age depending on which risk factors are present. A change in slope implies stretching or shrinking the scale on the age axis. A better way to interpret this is that the effect of the risk factors declines with age.

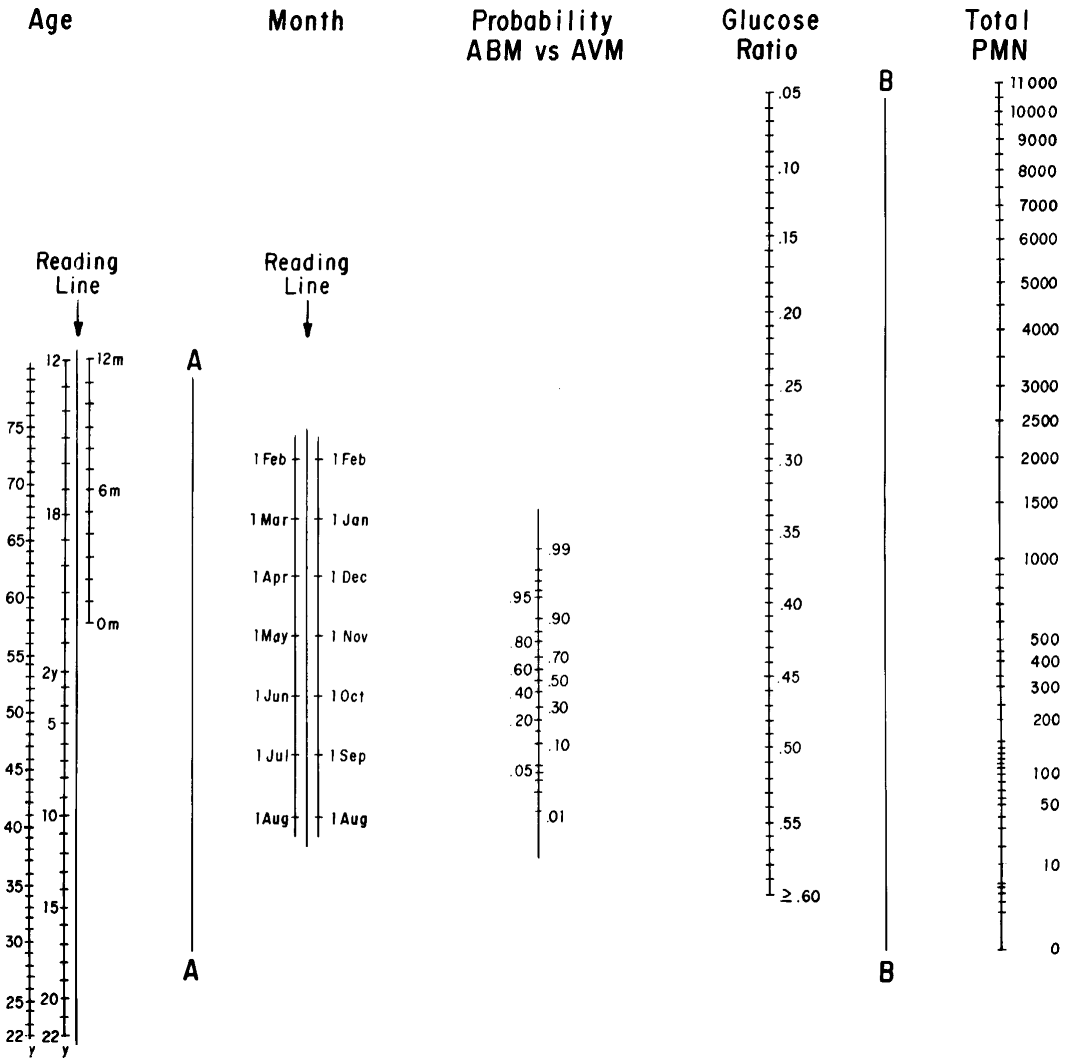


Figure 4: Nomogram for estimating probability of bacterial (ABM) vs. viral (AVM) meningitis. Depiction of a fitted binary logistic regression model. Step 1, place ruler on reading lines for patient’s age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) count in cerbrospinal fluid and mark intersection with line B; step 3, use ruler to join marks on lines A and B, then read off the probability of ABM vs. AVM [29].

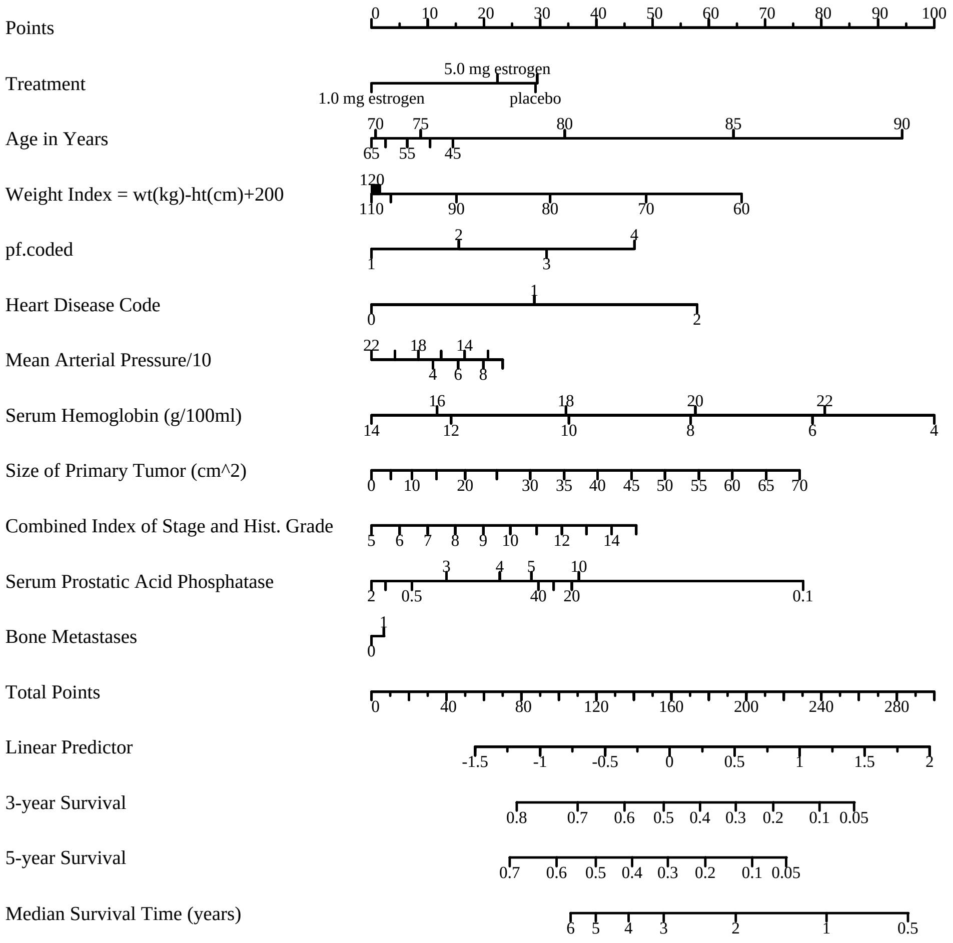


Figure 5: Software–generated nomogram for predicting death in a prostate cancer trial[4]. From a fitted Cox proportional hazards model predicting time until death (any cause). For each predictor one locates the value on that predictor's axis and then reads off the number of "severity points" on the top axis. These severity points are added manually and located on the "Total Points" axis. A vertical line drawn down from this value hits the 3–year survival probability, 5–year survival probability, and median survival time axes at the points corresponding to the predicted values. Median survival time stops at 6 years because patients were only followed up to 76 months.

References

[1] D. G. Altman and P. K. Andersen. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8:771–783, 1989.

[2] L. E. Braitman and F. Davidoff. Predicting clinical states in individual patients. *Annals of Internal Medicine*, 125:406–412, 1996.

[3] S. R. Brazer, F. S. Pancotto, T. T. Long III, F. E. Harrell, K. L. Lee, M. P. Tyor, and D. B. Pryor. Using ordinal logistic regression to estimate the likelihood of colorectal neoplasia. *Journal of Clinical Epidemiology*, 44:1263–1270, 1991.

[4] D. P. Byar and S. B. Green. The choice of treatment for cancer patients based on covariate information: application to prostate cancer. *Bulletin Cancer, Paris*, 67:477–488, 1980.

[5] D. Collett. *Modelling Binary Data*. Chapman and Hall, London, 1991.

[6] D. Collett. *Modelling survival data in medical research*. Chapman and Hall, London, 1994.

[7] J. Concato, A. R. Feinstein, and T. R. Holford. The risk of determining risk with multivariable models. *Annals of Internal Medicine*, 118:201–210, 1993.

[8] D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, 34:187–220, 1972.

[9] S. Derksen and H. J. Keselman. 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.

[10] S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8:551–561, 1989.

[11] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, 37:36–48, 1983.

[12] A. R. Feinstein. *Multivariable Analysis*. Yale University Press, New Haven, Connecticut, 1996.

[13] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247:2543–2546, 1982.

[14] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*, 3:143–152, 1984.

[15] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[16] F. E. Harrell, K. L. Lee, D. B. Matchar, and T. A. Reichert. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Cancer Treatment Reports*, 69:1071–1077, 1985.

[17] F. E. Harrell, K. L. Lee, and B. G. Pollock. Regression models in clinical studies: Determining relationships between predictors and response. *Journal of the National Cancer Institute*, 80:1198–1202, 1988.

[18] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 1989.

[19] W. A. Knaus, F. E. Harrell, C. J. Fisher, D. P. Wagner, S. M. Opan, J. C. Sadoff, E. A. Draper, C. A. Walawander, K. Conboy, and T. H. Grasela. 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.

[20] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, and D. P. Wagner. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122:191–203, 1995.

[21] A. Laupacis, N. Sekar, and I. G. Stiell. Clinical prediction rules: A review and suggested modifications of methodological standards. *Journal of the American Medical Association*, 277:488–494, 1997.

[22] K. L. Lee, D. B. Pryor, F. E. Harrell, R. M. Califf, V. S. Behar, W. L. Floyd, J. J. Morris, R. A. Waugh, R. E. Whalen, and R. A. Rosati. Predicting outcome in coronary disease: Statistical models versus expert clinicians. *American Journal of Medicine*, 80:553–560, 1986.

[23] G. Marshall, F. L. Grover, W. G. Henderson, and K. E. Hammermeister. Assessment of predictive models for binary outcomes: An empirical approach using operative death from cardiac surgery. *Statistics in Medicine*, 13:1501–1511, 1994.

[24] E. M. Ohman, P. W. Armstrong, R. H. Christenson, C. B. Granger, H. A. Katus, C. W. Hamm, M. A. O'Hannesian, G. S. Wagner, N. S. Kleiman, F. E. Harrell, R. M. Califf, E. J. Topol, K. L. Lee, and the GUSTO-IIa Investigators. Cardiac troponin T levels for risk stratification in acute myocardial ischemia. *New England Journal of Medicine*, 335:1333–1341, 1996.

[25] D. B. Pryor, F. E. Harrell, K. L. Lee, R. M. Califf, and R. A. Rosati. Estimating the likelihood of significant coronary artery disease. *American Journal of Medicine*, 75:771–780, 1983.

[26] D. B. Pryor, L. Shaw, F. E. Harrell, K. L. Lee, M. A. Hlatky, D. B. Mark, L. H. Muhlbaier, and R. M. Califf. Estimating the likelihood of severe coronary artery disease. *American Journal of Medicine*, 90:553–562, 1991.

[27] D. B. Pryor, L. Shaw, C. B. McCants, K. L. Lee, D. B. Mark, F. E. Harrell, L. H. Muhlbaier, and R. M. Califf. Value of the history and physical examination in identifying patients at increased risk for coronary artery disease. *Annals of Internal Medicine*, 118:81–90, 1993.

[28] L. R. Smith, F. E. Harrell, and L. H. Muhlbaier. Problems and potentials in modeling survival. In M. L. Grady and H. A. Schwartz, editors, *Medical Effectiveness Research Data Methods (Summary Report), AHCPR Pub. No. 92–0056*, pages 151–159. US Dept. of Health and Human Services, Agency for Health Care Policy and Research, Rockville, Maryland, 1992.

[29] A. Spanos, F. E. Harrell, and D. T. Durack. Differential diagnosis of acute meningitis: An analysis of the predictive value of initial observations. *Journal of the American Medical Association*, 262:2700–2707, 1989.

[30] D. J. Spiegelhalter. Probabilistic prediction in patient management. *Statistics in Medicine*, 5:421–433, 1986.

[31] J. C. van Houwelingen and S. le Cessie. Logistic regression, a review. *Statistica Neerlandica*, 42:215–232, 1988.