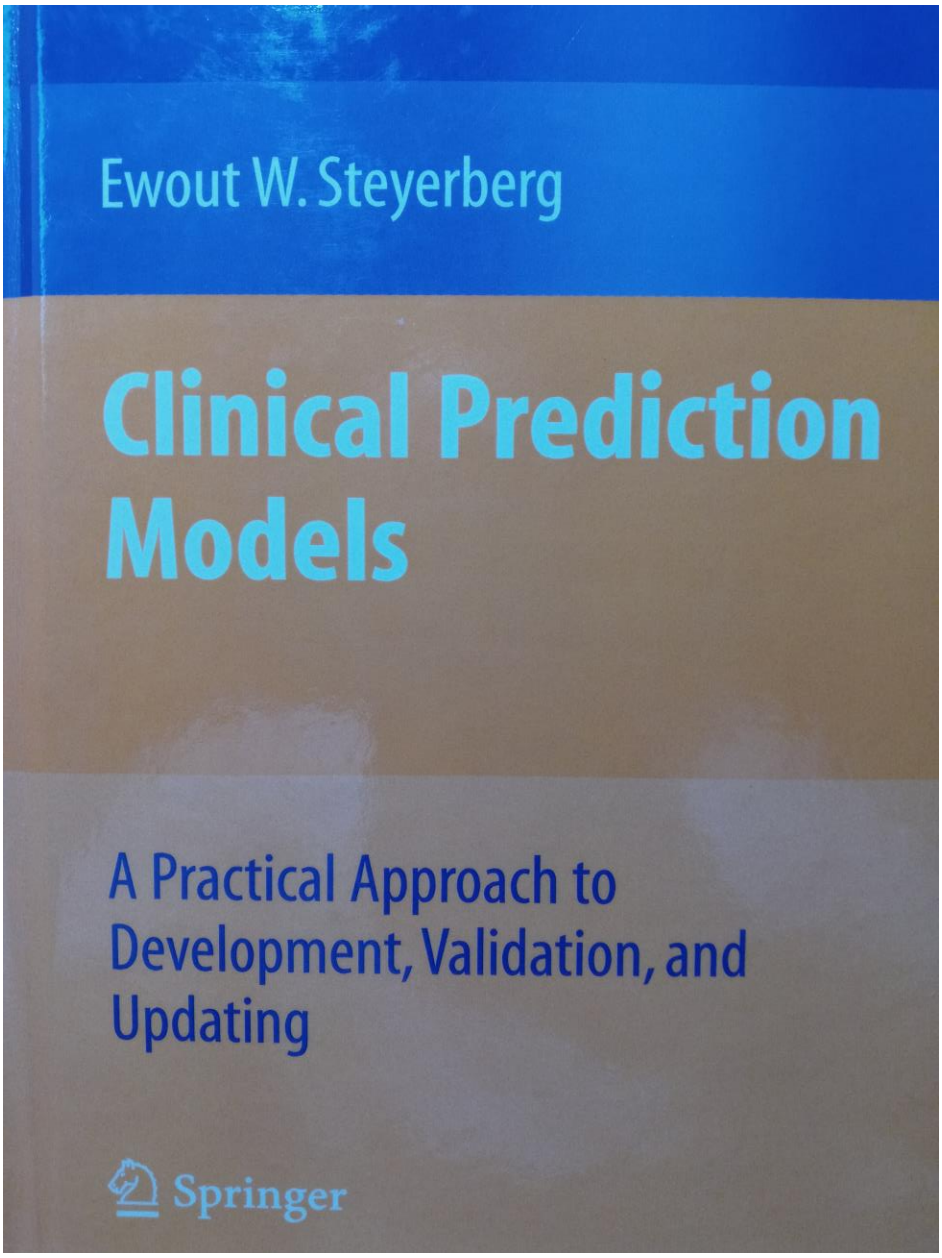
Ewout W. Steyerberg
Clinical Prediction Models
A Practical Approach to Development, Validation, and Updating
Springer

blood pressure, heart rate, infarct location, and age-by-Killip-class interaction. These predictors were previously found to comprise 90% of the predictive information of a more complex model for 30-day mortality in the GUSTO-I trial.[255] A review of RCTs published in the major medical journals after the year 2000 shows that covariate adjustment is used in approximately 50% of the cases.[339]

2.4.6 *Prediction Models and Observational Studies*

Confounding is the major concern in epidemiological analyses of observational studies. When treatments are compared, groups are often quite different because of a lack of randomization. Subjects with specific characteristics are more likely to have received a certain treatment than other subjects ("indication bias", Fig. 2.7). If these characteristics also affect the outcome, a direct comparison of treatments is biased, and may merely reflect the lack of initial comparability ("confounding"). Instead of treatment, many other factors can be investigated for their causal effects. Often, randomization is not possible, and observational studies are the only possible design. Dealing with confounding is an essential step in such analyses.

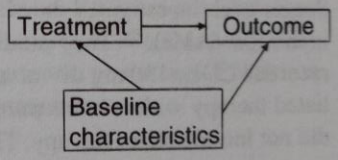

Fig. 2.7 Schematic representation of confounding in an observational study. Baseline characteristics act as confounders since they are related to the treatment and to the outcome

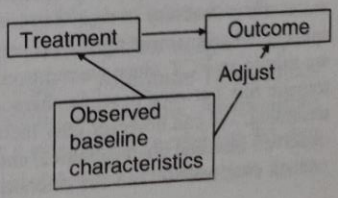

Fig. 2.8 Schematic representation of adjustment for baseline characteristics in an observational study. By adjustment, we aim to correct for the systematic link between observed baseline characteristics and outcome, hence answering the question what the treatment effect would be if observed baseline characteristics were similar between treatment groups

*4.1.1 Examples of Linear Regression

An example of a medical outcome is blood pressure. We may want to predict the blood pressure after treatment with an anti-hypertensive or other intervention.[87] Also, quality of life scales may be relevant to evaluate.[242] Such scales are strictly speaking only ordinal, but can for practical purposes often be treated as continuous outcomes. A specific issue is that quality of life scores have ceiling effects, because minimum and maximum scores apply.

4.1.2 Economic Outcomes

Health economics is another important field where continuous outcomes are considered, such as length of stay in hospital, or length of stay at a specific ward (e.g. the intensive care unit), or total costs for patients.[88]

Cost data are usually not normally distributed. Such economic data have special characteristics, such as patients without any costs (zero), and a long tail because some patients having considerable costs. We might consider the median as a good descriptor of the outcome. Interestingly, we are however always interested in the mean costs, since the expectation is what matters most from an economical perspective. Sometimes analyses have been performed to identify "high-cost" patients, after dichotomizing the outcome at some cost threshold.

*4.1.3 Example: Prediction of Costs

Many children in moderate climates suffer from an infection by the respiratory syncytial virus (RSV). Some children, especially premature children are at risk of a severe infection, leading to hospitalization. The mean RSV hospitalization costs were 3,110 euros in a cohort of 3,458 infants and young children hospitalized for severe RSV disease during the RSV seasons 1996–1997 to 1999–2000 in the Southwest of The Netherlands. RSV hospitalization costs were higher for some patient categories, e.g. those with lower gestational age or lower birth weight, and younger age. The linear regression model had an adjusted R^2 of 8%.[345] This indicates a low explanatory ability for predicting hospitalization costs of individual children. However, the model could accurately estimate the anticipated mean hospitalization costs of groups of children with the same characteristics. These predicted costs were used in decision analyses of preventive strategies for severe RSV disease.[46]

4.1.4 Transforming the Outcome

An important issue in linear regression is whether we should transform the outcome variable. The residuals ($y-\hat{y}$) from a linear regression should have a normal distribution with a constant spread ("homoscedasticity"). This can sometimes be achieved by

e.g. a log transformation for cost data, but other transformations are also possible. As Harrell points out, transformations of the outcome may reduce the need to include transformations of predictor variables.[174] Care should be taken in backtransforming predicted mean outcomes to the original scale. Predicted medians and other quantiles are not affected by transformation. The log–normal distribution can be used for the mean on the original scale after a log transformation, but a more general, non-parametric, approach is to use "smearing" estimators.[341]

4.1.5 *Performance: Explained Variation*

In linear regression analysis, the total variance in y ("total sum of squares", TSS) is the sum of variability explained by one or more predictors ("model sum of squares", MSS) and the error ("residual sum of squares", RSS):

$$\text{TSS} = \text{MSS} + \text{RSS}$$
$$\text{var (regression on } x_i\text{)} + \text{var (error)} = \sum (\hat{y} - \text{mean } (y))^2 + \sum (y - \hat{y})^2$$

The estimates of the variance follow from the statistical fit of the model to the data, which is based on the analytical solution of a least squares formula. This fit minimizes the error term in the model, and maximizes the variance explained by *xi*. Better prediction models explain more of the variance in y. R^2 is defined as MSS / TSS.[472]

To appreciate values of R^2, we consider six hypothetical situations where we predict a continuous outcome y, which has a standard normal distribution (N(0,1), i.e. mean 0 and standard deviation 1) with one predictor x (N(0,1)). The regression coefficients for x are varied in simulations, such that R^2 is 95%, 50%, 20%, 10%, 5%, and 0% (Fig. 4.1). We note that an R^2 of 95% implies that observed outcomes are always very close to the predicted values, while gradually relatively more error occurs with lower R^2 values. When R^2 is 0%, no association is present.

To appreciate R^2 further, we plot the distributions of predicted values ($\hat{y}$). The distribution of $\hat{y}$ is wide when R^2 is 95%, and very small when R^2 is 5%, and near a single line when R^2 is 0% (Fig. 4.2). The distribution of y is always normal with mean 0 and standard deviation 1.

4.1.6 *More Flexible Approaches*

The generalized additive model (GAM) is a more flexible variant of the linear regression model.[180, 181, 472] A GAM allows for more flexibility especially for continuous predictors. It replaces the usual linear combination of continuous predictors with a sum of smooth functions to capture potential non-linear effects: $y = b_0 + f_i(x_i)$ + error, where f_i refers to functions for each predictor, e.g. loess smoothers.