

## **Chapter 2**

### **Validity of prediction models: when is a model clinically useful?**

#### **Abstract**

Prediction models combine patient characteristics to predict medical outcomes. Unfortunately, such models do not always perform well for other patients than those from whose data the models were derived. Therefore, the validity of prediction models needs to be assessed in new patients. Predicted probabilities can be calculated with the model and compared with the actually observed outcomes. We may distinguish several aspects of validity: i) agreement between observed probabilities and predicted probabilities (calibration), ii) ability of the model to distinguish subjects with different outcomes (discrimination), iii) ability of the model to improve the decision making process (clinical usefulness). We discuss those aspects and illustrate some measures using models for testicular germ cell and prostate cancer. We conclude that good calibration and discriminative ability are not sufficient for a model to be clinically useful. Application of a prediction model is sensible, if the model is able to provide useful additional information for clinical decision making.

#### **Introduction**

Prediction models have been developed to predict the outcome of future patients in many clinical fields. These outcomes are often events during the patient's disease process. Examples are the diagnosis of a residual mass as 'benign' or 'tumour', or the prediction of recurrence of disease up to a specified time point. Modelling techniques like regression analysis and neural networks relate the outcomes to patient characteristics, also labelled predictors.<sup>1-3</sup> The resulting model may divide patients into subgroups (e.g., high/intermediate/low risk for relapse), or predict individual probabilities for the outcome (e.g. 'the probability that a mass is benign is 80%'). Here, we focus on the predicted probabilities for individual patients.

A model can be presented in several formats. Sometimes the formula for the linear predictor is presented. Such a formula contains the exact regression coefficients and the functions to transform the predictor variables. Often a more user-friendly format like a score chart or a nomogram is preferable. A score chart relates scores to values of the predictors. If the model was developed with regression analysis, the scores are directly derived from the regression coefficients, often by multiplying by 10 and rounding. The sum of the scores plus

a constant is the sumscore. The sumscore is easy to calculate, in contrast to the linear predictor. Both are mathematically related to the predicted probability. For the development of nomograms, the same principle as for score charts can be applied.

A prediction model usually performs better in the sample used to develop the model (development sample) than in other samples, even if those samples are derived from the same population.<sup>1,4</sup> This ‘optimism’ is most marked when the development sample is small. Several approaches have been proposed to assess the performance in samples of the same population as the development sample (internal validation). A typical technique for internal validation is cross-validation. With this technique, the model is developed on a randomly drawn part of the development sample and tested on the rest of the sample. This is repeated several times and the average is taken as an estimate of performance. Another technique is bootstrapping. A sample of the same size as the development sample is randomly drawn from that sample with replacement. Models are then developed in the bootstrap samples and tested in the original sample or in those subjects not included in the bootstrap samples.<sup>5,6</sup> In addition, the performance of prediction models can be determined in samples from different but related populations, like patients from other centres or treated in more recent years. The performance in such samples is labelled external validity or ‘transportability’.<sup>7,8</sup>

Irrespective of the type of validation study (internal or external), several aspects of validity can be distinguished.<sup>9</sup> Calibration refers to the agreement between observed probabilities and predicted probabilities. For instance, if a 60% probability is predicted that the event will occur, then the event should on average occur in every 60 out of 100 patients with predicted probabilities of 60%. However, a model that reliably predicts probabilities that are all between 40% and 60% is not able to distinguish patients with the outcome from patients without the outcome. The discriminative ability of such a model is poor. Ideally, predicted probabilities are close to 0% or 100%. A third aspect, clinical usefulness, indicates whether the model can be helpful to a clinician in making a decision, such as order a test, start a certain treatment or perform surgery. It may occur that calibration and discriminative ability of the model are satisfactory for a group of patients, but that the decisions based on the model are the same as the decisions based on the reference policy. This may particularly be so, if the reference policy already considers clinical information which is also included in the prediction model. In that case, the model has no additional value and is not clinically useful.

We discuss these three aspects of model validity in the next section. For each aspect we give several measures (Table 2.1). The measures are illustrated with a model predicting the histology of a residual mass in testicular germ cell cancer patients, which is a binary outcome (Case study 1). In addition, we discuss the same measures for a model developed with survival data (Case study 2). This model predicts the probability of a patient treated for clinically localised prostate cancer to be free of recurrence after 7 years of follow-up.

**Table 2.1** Measures for three aspects of model validity

| <i>Aspects</i>                                    | <i>Measures</i>  | <i>Characteristics</i>  |
|---|--|---|
| Calibration                                       | Calibration plot   | Visual impression of observed frequencies vs. predicted probabilities   |
|   | - slope  | Estimate of extremeness of predicted probabilities  |
|   | -intercept/<br>'calibration-in-the-large'<br>- $E_{avg}$ | Estimate of systematically too high/low predicted probabilities<br>Average absolute difference between observed frequencies and predicted probabilities |
|   | Hosmer-Lemeshow statistic                                | Test for 'goodness-of-fit', i.e. deviance of grouped observed outcomes and predicted outcomes   |
| Discrimination                                    | Boxplot of predicted probabilities                       | Visual impression of spread in predicted probabilities; relies on adequate calibration  |
|   | c-statistic (ROC area)                                   | Summary of quality over a range of threshold values   |
| Clinical usefulness<br>(threshold value required) | Accuracy   | Percentage of patients correctly classified, given a certain threshold value  |
|   | Sensitivity  | Percentage of patients with the outcome correctly classified as diseased  |
|   | Specificity  | Percentage of patients without the outcome correctly classified as non-diseased   |
|   | Decrease in weighed false classifications                | Model and reference policy are compared by weighing patients falsely classified as diseased and non-diseased according to relative severity             |

## **Case study 1**

### *Clinical problem*

Retroperitoneal metastatic disease in patients with nonseminomatous testicular germ cell cancer is generally successfully treated with cis-platin based chemotherapy. Residual masses after treatment may be totally benign or may still contain tumour elements. If it is suspected that the residual mass contains tumour elements, the mass should be resected. Frequent observation during follow-up may be an alternative approach, if the risk of tumour is low. We developed a prediction model to predict the probability that the histology of a residual mass is totally benign.<sup>10</sup> Since only two values for the histology of a residual mass are possible in our model: 'yes, totally benign' or 'no, not totally benign', the outcome variable is binary and the model was developed with logistic regression analysis. Predictors for benign tissue were a teratoma free primary tumour, normal levels before chemotherapy of serum tumour markers (alpha-fetoprotein [AFP] and human chorionic gonadotropin [HCG]),

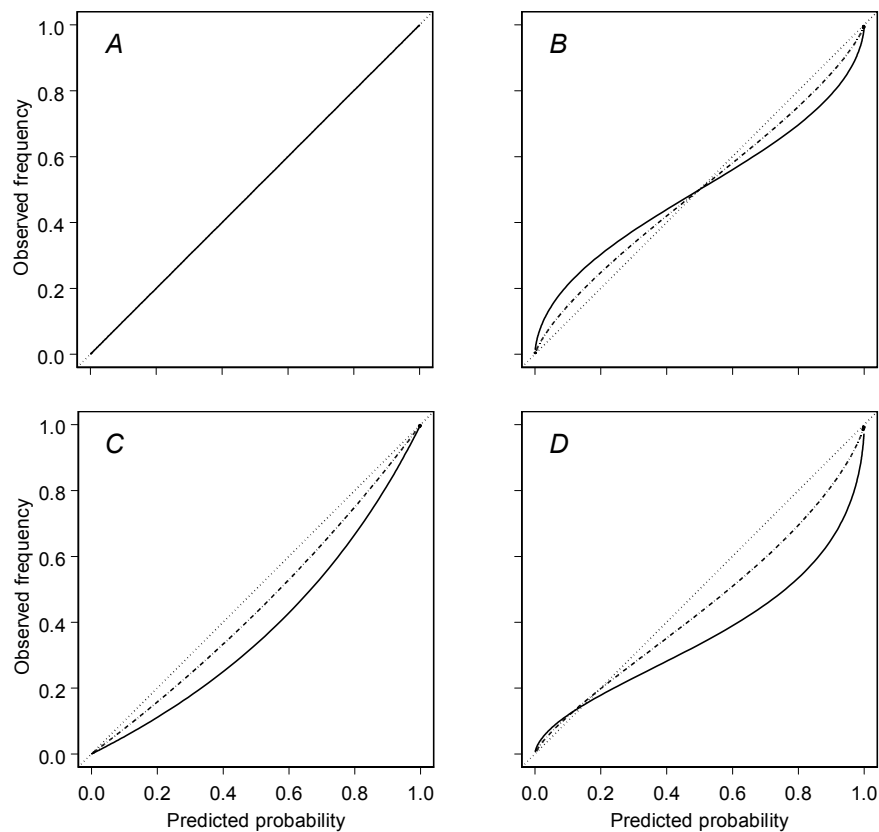
a small residual mass after chemotherapy, and a large reduction in mass size during chemotherapy. Mass size and change in mass size were divided in 3 categories (< 5 cm; 5-10 cm [MASS1]; > 10 cm [MASS2] and no change; reduction; progression). The linear predictor can be calculated as  $-1.56 + 0.99*TER + 0.93*AFP + 0.58*HCG - 1.05*MASS1 - 0.54*MASS2 + 0.87*REDUC - 2.03*PROG$ , where the variables have value 1 or 0. The predicted probability of benign tissue can be calculated with the formula  $1/(1+\exp[-\text{linear predictor}])$ .

To study the external validity of the model, we calculated the predicted probabilities of benign tissue for 276 patients from Indiana University Medical Center (IUMC) and compared the actual observed histologies with the predicted probabilities.<sup>11</sup> 76 patients had benign tissue and 200 patients had masses containing tumour elements at surgery.

### *Calibration*

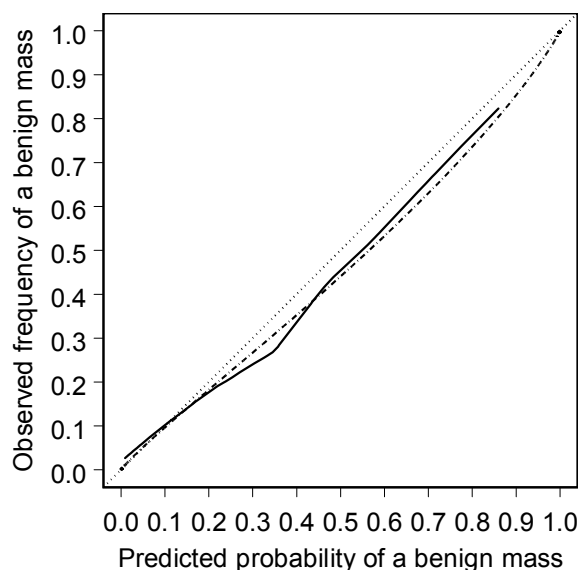
An important question to decide whether a model is valid, is ‘Are the predictions of the model reliable?’. To answer this question, the predicted probabilities as calculated with the model need to be compared with the observed probabilities. However, we only observe that a residual mass is benign or not, which has a value of 0 (for tumour) or 1 (for benign) not a probability between 0 and 1. To estimate an observed probability for each patient, a smoothing technique can be applied or patients can be grouped. Figure 2.1 shows smoothed lines through all 0 and 1 values (y-axis) of patients with different predicted probabilities between 0 and 1 (x-axis). The observed outcome values 0 and 1 are replaced with values between 0 and 1 by combining outcome values of patients with similar predicted probabilities.<sup>12</sup> Those values may be considered as ‘observed probabilities’. Since the observed probability remains a theoretical concept, we will use the term observed frequency.

A plot in which observed frequencies are plotted against predicted probabilities is called a calibration plot.<sup>13</sup> Ideally, if observed frequencies and predicted probabilities agree over the whole range of probabilities, the plot shows a 45° line (Figure 2.1A). The intercept of the line ( $\alpha$ ) is 0 and the slope ( $\beta$ ) is 1. The intercept and slope of the calibration line can be estimated in a logistic regression model with the linear predictor, calculated for the patients of the validation data, as the only predictor variable:  $\log \text{odds (observed histology is benign)} = \alpha + \beta \text{ linear predictor}$ .<sup>14</sup> The observed histology is 0 or 1; the linear predictor is calculated as mentioned before. A slope smaller than 1 indicates optimism (Figure 2.1B). This means that the predictions are too extreme: too low estimates for low predictions and too high estimates for high probabilities. The opposite, a slope larger than 1, indicates that the predicted probabilities are not sufficiently extreme.<sup>1,15</sup> An intercept different from 0 indicates that the predicted probabilities are systematically too high (intercept < 0, Figure 2.1C) or too low (intercept > 0). If both the slope differs from 1 and the intercept differs from 0, the interpretation of the miscalibration is difficult, because the values of intercept and slope are related (Figure 2.1D).



**Figure 2.1** Theoretical calibration plots to study the agreement between observed frequencies and predicted probabilities with a smoothed line through all outcome values (0 and 1). Figure A shows agreement between observed frequencies and predicted probabilities (intercept=0; slope=1). Figure B shows optimistic predicted probabilities: predictions are too extreme (intercept=0; slope=0.8 [---] or 0.6 [—]). Figure C shows too high predicted probabilities (intercept=log(0.75)[---], or log(0.5) [—], slope=1). Figure D shows a combination of optimism and ‘miscalibration-in-the-large’ (intercept=log(0.75) or log(0.5), slope=0.8 or 0.6 for --- and — respectively).

Figure 2.2 shows the calibration plot of the model predicting the residual mass histology for the 276 patients from IUMC. The slope was close to 1 (0.91; 95% confidence interval: 0.64 – 1.18), which indicates that the model is not too optimistic. Since the value of the intercept is related to the value of the slope, the intercept automatically changes when the slope changes. Therefore, the slope was fixed at 1 to make a sensible interpretation of the value of the intercept. In the IUMC data the intercept was -0.19 (95% confidence interval: -0.50 – 0.11). The exponent of the intercept in a logistic regression model is the observed odds of the outcome divided by the predicted odds, which was 0.82 ( $\exp[-0.19]$ ). This indicates that the predicted odds were on average 20% ( $1/0.82 = 1.2$ ) too high and the predicted probabilities were somewhat less than 20% too high.



**Figure 2.2** Calibration plot of the IUMC data. Parametric (---) and non-parametric (—) lines were created by regression analysis with the linear predictor as covariate or with a smoothing technique. The lines are in general below the ideal line (....). The intercept of the calibration curve is - 0.19, when the slope is fixed at 1. The average absolute distance between the ideal and parametric line is  $E_{avg}$ .

This type of miscalibration with the slope close to 1 and the intercept different from 0, is a typical finding in external validation studies. It indicates that certain patient characteristics, which were not included in the prediction model, were differently distributed in the validation sample compared with the development sample. For example, patients of the validation sample were derived from a tertiary referral hospital, while patients from the development sample were mainly derived from secondary referral hospitals.

The overall difference between observed frequencies and predicted probabilities can also be studied by comparing the sum of all predicted probabilities with the number of observed outcomes ('calibration-in-the-large'). The sum of all predicted probabilities for benign tissue was 84 (31%) for the patients of IUMC, which was higher than the observed 76 (28%) patients with benign tissue. This result also shows that the predicted probabilities were in general too high.

Another appealing measure is the absolute difference between the observed frequencies and the predicted probabilities ( $E$ ). In the calibration plot, the absolute difference between those two was on average 3.6%.

Observed frequencies can also be determined by calculating the proportion of totally benign tissue for groups of patients. Patients may, for instance, be grouped by their predicted probabilities. The calculated proportion of outcomes serves then as observed frequency for patients of the relevant group. Agreement between those observed frequencies and predicted probabilities can be statistically tested with the Hosmer-Lemeshow (H-L) goodness-of-fit test.<sup>2</sup> The test statistic sums up the differences between the observed frequencies and

predicted probabilities of each group. The statistic follows a chi square distribution. For external validation, the number of degrees of freedom equals the number of groups. A high H-L statistic is related to a small p-value and indicates lack of fit.

**Table 2.2** Comparison of observed frequencies and predicted probabilities in grouped patients with the Hosmer-Lemeshow statistic

|   | <i>N</i> | <i>Predicted</i> | <i>Observed</i> | <i>Contribution<br/>H-L statistic</i> |
|---|----------|------------------|-----------------|---------------------------------------|
| <i>Equal intervals</i>                                |          |                  |                 |                                       |
| 0 – 10%   | 54       | 4.0%             | 5.6%            | 0.37                                  |
| 11 – 20%  | 59       | 14.5%            | 15.3%           | 0.03                                  |
| 21 – 30%  | 25       | 24.1%            | 12.0%           | 2.00                                  |
| 31 – 40%  | 61       | 34.0%            | 26.2%           | 1.63                                  |
| 41 – 50%  | 12       | 47.0%            | 33.3%           | 0.90                                  |
| 51 – 60%  | 33       | 57.0%            | 57.6%           | 0.01                                  |
| 61 – 70%  | 8        | 67.8%            | 62.5%           | 0.10                                  |
| 71 – 80%  | 16       | 74.7%            | 62.5%           | 1.26                                  |
| 81 – 90%  | 8        | 86.1%            | 87.5%           | 0.01                                  |
| 91 – 100%   | 0        | -                | -               | -                                     |
|   |          |                  |                 | total: 6.3, p = 0.71                  |
| <i>Equal size</i>                                     |          |                  |                 |                                       |
| 0 – 2.8%  | 27       | 1.9%             | 3.7%            | 0.45                                  |
| 2.9 – 6.9%  | 27       | 6.0%             | 7.4%            | 0.10                                  |
| 7.0 – 11.8%   | 20       | 11.0%            | 5.0%            | 0.74                                  |
| 11.9 – 17.4%  | 34       | 16.0%            | 17.6%           | 0.06                                  |
| 17.5 – 24.9%  | 25       | 22.5%            | 16.0%           | 0.61                                  |
| 25.0 – 32.2%  | 17       | 30.0%            | 41.2%           | 1.01                                  |
| 32.3 – 35.9%  | 36       | 33.8%            | 16.7%           | 4.73                                  |
| 36.0 – 50.5%  | 27       | 42.2%            | 33.3%           | 0.86                                  |
| 50.6 – 58.6%  | 29       | 57.3%            | 55.2%           | 0.05                                  |
| 58.7 – 100%   | 34       | 74.8%            | 70.6%           | 0.33                                  |
|   |          |                  |                 | total: 8.9, p = 0.54                  |
| <i>Grouped by co-<br/>variate values <sup>a</sup></i> |          |                  |                 |                                       |
| ter -; < 5 cm   | 63       | 61.0%            | 55.6%           | 0.77                                  |
| ter -; 5 – 10 cm                                      | 26       | 28.8%            | 42.3%           | 2.30                                  |
| ter -; > 10 cm  | 16       | 33.1%            | 37.5%           | 0.14                                  |
| ter +; < 5 cm   | 58       | 35.2%            | 20.7%           | 5.34                                  |
| ter +; 5 – 10 cm                                      | 56       | 14.1%            | 14.3%           | 0.001                                 |
| ter +; > 10 cm  | 57       | 11.6%            | 7.0%            | 1.16                                  |
|   |          |                  |                 | total: 9.7, p = 0.14                  |

<sup>a</sup> ter -: primary tumour teratoma-negative; ter +: teratoma-positive

Several grouping strategies can be used to estimate the H-L statistic. Groups can be of equal sizes (e.g. percentiles of predicted probabilities) or be of equal prediction-intervals (e.g. 0 – 10%; 10 – 20%; ..; 90 – 100%). Most statistical packages use percentiles of groups to calculate the H-L statistic, because grouping by intervals may result in groups with very few or no patients at all. Unfortunately, the H-L statistic has little power to discover miscalibration in small samples. In contrast, even small disagreements between observed frequencies and predicted probabilities result in a significant H-L statistic in large samples.<sup>16</sup>

Table 2.2 shows the results of alternative grouping strategies for the IUMC data (n=276). Grouping by equal prediction intervals resulted in a few groups with many patients (e.g. 61 patients had a predicted probability between 31 and 40%) and one group with no patients at all (interval 91 – 100%). Grouping by equal size did not result in groups with exactly the same number of patients per group (17 to 36 patients per group), because some patients had the same predicted probability and were assigned to the same percentile. Both grouping strategies resulted in non-significant H-L statistics (p-values 0.71 and 0.54). Thus in contrast to the calibration plot, the H-L statistic did not indicate disagreement between observed frequencies and predicted probabilities for the patients of IUMC.

A concern with grouping by predicted probabilities is, that groups may contain patients with widely different values of the covariates indicating that the patients are not similar at all. Therefore, we also grouped the patients by covariate patterns. We used the values of the covariates ‘histology of primary tumour’ and ‘size of residual mass’. This resulted in a different p-value than grouping by predicted probabilities, although none of the p-values were significant. However, it can easily be shown that the different grouping strategies may result in different implications. Thus, grouping of patients is necessary to estimate the H-L statistic, but it is also the weakness, since the result is sensitive to the choice of groups.

### *Discriminative ability*

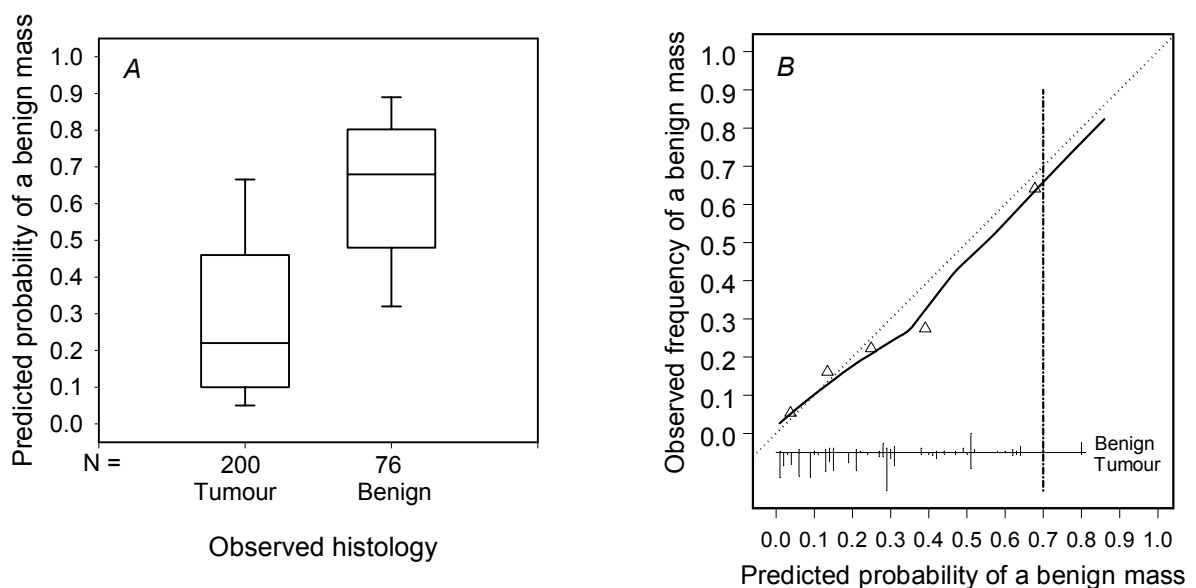
A prediction model can excellently distinguish patients with different outcomes, if it predicts probabilities close to 100% for patients with the outcome and probabilities close to 0% for patients without the outcome. Thus, a good discriminating model will show a wide spread in the distribution of the predicted probabilities, away from the average probability. Figure 2.3 shows the distribution of the probabilities for patients from IUMC per outcome value (benign or tumour). Patients with masses containing tumour elements have in general low predicted probabilities for benign tissue (lower than 33% for 75% of the patients). However, the probabilities of patients with benign tissue should be closer to 1 for better discriminative performance.

The distribution of the predicted probabilities can also be indicated in the calibration plot. We prefer then to speak of a ‘validation plot’. Percentiles of predicted probabilities can be indicated by symbols along the calibration line (triangles in Figure 2.3B). Symbols close together indicate limited discriminative ability. Furthermore, the distribution of predicted



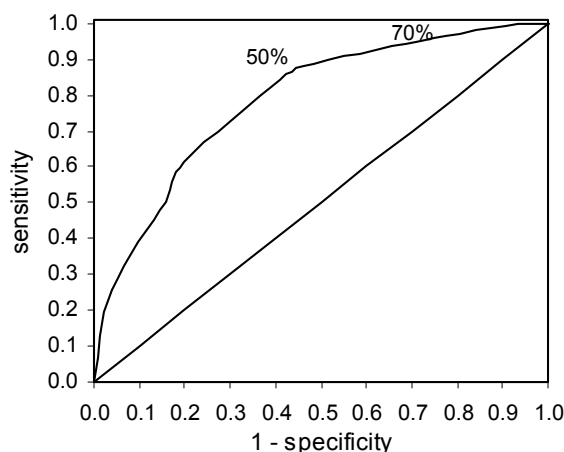
probabilities can be shown per outcome value at the bottom of the plot. In figure 2.3B, lines upwards represent patients with benign tissue, lines downwards patients with tumour.

The concordance-statistic ( $c$ -statistic) is often used to quantify discriminative ability. For binary outcomes the  $c$ -statistic is identical to the area under the receiver operating characteristic curve (ROC curve). The ROC curve is a plot of true-positive rate (e.g. percentage of patients having the outcome and correctly classified as diseased, or sensitivity) versus false-positive rate (e.g. percentage of patients having the outcome and incorrectly classified as diseased, or 1-specificity) evaluated at consecutive threshold values of the predicted probability (Figure 2.4). The area under the ROC curve represents the probability that a patient with the outcome has a higher predicted probability than a patient without the outcome for a random pair of patients consisting of one patient with and one patient without the outcome. Pairs of patients with the same outcome are not considered. A useless prediction model, such as a coinflip, would realise an area of 50%. When the area is 100%, the model discriminates perfectly.<sup>1,17</sup>



**Figure 2.3** Boxplot (A) showing the distributions of the predicted probabilities per outcome value for the IUMC data: graphical impression of discrimination. The boxes show the 25%, 50% and 75% cumulative frequencies. The ends of the lines show the 10% and 90% cumulative frequencies.

Validation plot for the IUMC data (B). Calibration is shown with a non-parametric line (—). Distribution of the predicted probabilities is indicated by groups of patients (triangles) and for individual patients (vertical lines). Vertical lines upwards represent patients with benign tissue, lines downwards represent tumour. At the threshold value of 70% (---) only a few patients have predicted probabilities above the threshold value.



**Figure 2.4** Receiver operating characteristic (ROC) curve for the IUMC data. The area under the curve is 0.79. The threshold values 50% and 70% are indicated.

Since sensitivity usually reflects the proportion of patients correctly classified as having the disease, we calculated the sensitivity of the prediction model as the proportion of patients with tumour elements in their masses correctly classified as tumour, e.g. with predicted probabilities of benign tissue equal or lower than the threshold value. Specificity is then the proportion of patients having complete benign masses with predicted probabilities of benign tissue above the threshold value. For the patients of IUMC we calculated an area under the curve of 0.79 (Figure 2.4).

### *Clinical usefulness*

In general, measures of calibration and discriminative ability evaluate the performance of a prediction model over the whole range of predicted probabilities. This is statistically attractive, but assumes that all predicted probabilities are equally relevant. However, in our case it is not very much of interest whether a predicted probability of 10% actually should be 20%, if only patients with relatively high predicted probabilities for benign tissue will not be operated. More important is to define a clinically relevant threshold value, e.g. the probability of benign tissue below which the expected benefits of surgery outweigh the expected surgical risks (morbidity and mortality).<sup>18</sup> The expected benefits and costs of the treatment options (surgery yes/no) have been weighted in a decision analysis. A threshold value of 70% seems sensible; predicted probabilities below 70% indicate surgery and probabilities above 70% indicate no surgery.<sup>19</sup>

Hence, clinically relevant information is the proportion of patients that receive the correct treatment if the candidates for operation are selected by the prediction model with a threshold value of 70%. This proportion is well known as the accuracy (proportion of patients correctly classified either as having benign tissue or tumour). Proportions of

correctly classified patients may also be determined given the outcome value (sensitivity and specificity).

Table 2.3 shows that 76% of all patients from IUMC were correctly classified. Nearly all patients with tumour elements in their masses had predicted probabilities under 70% (sens=97%, 193/200). However, only a few patients had benign tissue with predicted probability for benign tissue above 70% (spec=22%, 17/76).

The low specificity shows that most of the patients with a benign mass still would be operated. Compared with the current policy (all patients were operated), application of the model would only spare 9% (24/276) of the patients an operation. Therefore, one may doubt the clinical usefulness of the model.

A threshold value of 70% implies that leaving a tumour unresected (false-negative classification) is considered 7/3 (= 2.3) times worse than resecting a benign mass (false-positive classification).<sup>20</sup> This ratio can be used to weigh false classifications. We can express the number of tumours left unresected as the number of unnecessarily resected masses (URM), i.e. 1 tumour left unresected equals 7/3 URMs. If all masses were resected (reference policy), 76 times a benign mass would be unnecessarily resected ( $76 \times 1 = 76$  URM). Resecting the masses with predicted probability below 70% would result in 59 unnecessarily resected masses with benign tissue (Table 2.4) and 7 tumour masses would incorrectly not be resected (in total:  $59 \times 1 + 7 \times 7/3 = 75$  URM). The clinical usefulness of the model can be quantified as the relative reduction in URM compared to the reference policy ( $[(76-75)/76=1\%]$ ). We can conclude that the clinical usefulness of the prediction model is marginal for the patients who are currently considered for resection at IUMC.

An impression of clinical usefulness can also be achieved with a validation plot. Figure 2.3B shows that only a few predicted probabilities were above 70%. Even if the model would be recalibrated (i.e. adjustment of the intercept of the prediction model) to improve the reliability of the predictions, only a few patients would not be operated.

**Table 2.3** Measures related to clinical usefulness estimated for the IUMC and Cleveland data

| <i>Measure</i>                              | <i>IUMC data</i> | <i>Cleveland data</i> |
|---|------------------|-----------------------|
| Accuracy                                    | 76%              | 81%                   |
| Sensitivity                                 | 97%              | 30%                   |
| Specificity                                 | 22%              | 99%                   |
| Decrease in URM/UPD                         | 1%               | 26%                   |
| URM: unnecessarily resected masses          |                  |                       |
| UPD: untreated while progression of disease |                  |                       |

## Case study 2

### *Clinical problem*

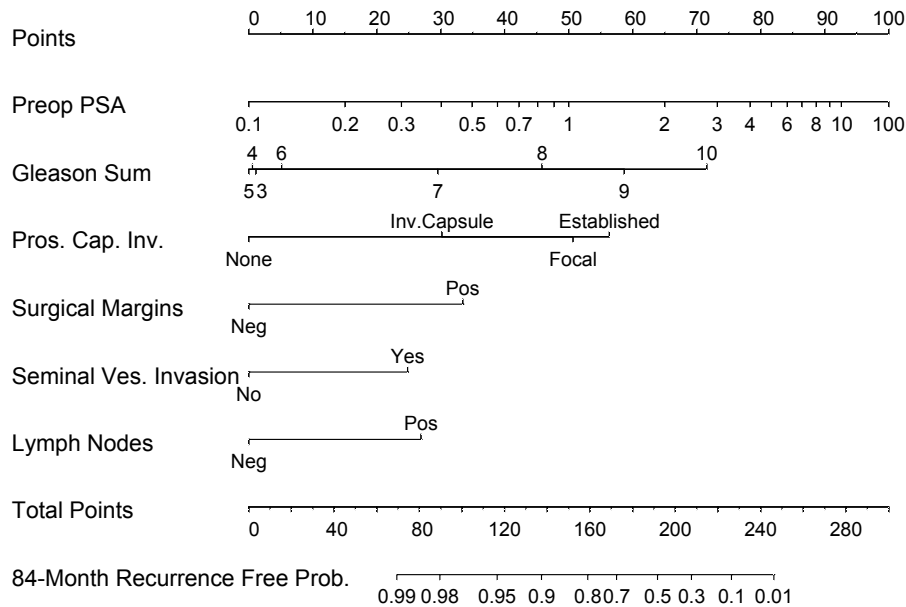
Approximately one third of men treated with radical prostatectomy for clinically localised prostate cancer later experience progression of their disease. Early identification of men likely to experience disease progression would be useful in considering early adjuvant therapy. Kattan *et al.* modelled the 7-year probability of disease progression using data of 996 men treated with radical prostatectomy at one single centre.<sup>21</sup> The outcome variable of this model is recurrence within the first 7 years after prostatectomy. This is again a binary variable (recurrence yes/no), however not all patients are followed for the complete 7 years. Survival data contain information about the event (does the patient experience disease progression yes/no) and about the time of event, or survival time (how many months is the patient free of progression). In our case study, survival time is initiated at prostatectomy and extends until the patient has recurrence of disease. If the patient is lost to follow-up, the survival time is censored at the last follow-up time.

In order to be able to include patients with a limited follow-up time in the analysis, Cox proportional hazard regression analysis was used for the model development. The predictors were preoperative serum PSA level, Gleason sum in the surgical specimen, prostatic capsular invasion, surgical margin status, seminal vesicle invasion and lymph node status. The model was presented as a nomogram (Baylor nomogram, Figure 2.5).<sup>21</sup>

The external validity of the nomogram was studied in 946 patients from the Cleveland Clinic, Ohio with a median follow-up time of 34 months. During follow-up, 144 patients had progression of disease and 38 patients were still at risk at 7 years after prostatectomy.

### *Calibration*

The nomogram for prostate cancer predicts the probability that a patient is free of recurrence after 7 years of follow-up. To study calibration, the predicted probability might hence be compared with the observed frequency that the patient is free of recurrence after 7 years. Determining the observed frequency of recurrence by calculating the frequency of observed recurrences in the data would give an underestimation of the probability. A substantial part of the patients was not followed the complete seven years and would have relapsed, if they were followed the complete period. Relevant observed frequencies can be estimated with the Kaplan-Meier method, which is a standard method to deal with censoring, i.e. incomplete follow-up.



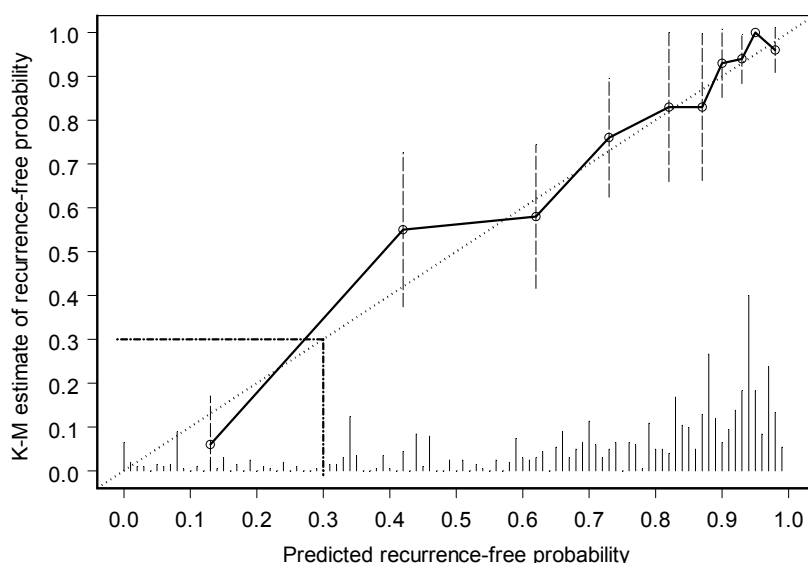
**Figure 2.5** Postoperative nomogram for predicting recurrence after radical prostatectomy. Reprinted with permission.<sup>21</sup>

**Instructions for Physician:** Locate the patient's PSA on the PSA axis. Draw a line straight upwards to the Points axis to determine how many points towards recurrence the patient receives for his PSA. Repeat this process for the other axes, each time drawing straight upward to the Points axis. Sum the points achieved for each predictor and locate this sum on the Total Points axis. Draw a line straight down to find the patient's probability of remaining recurrence free for 84 months assuming he does not die of another cause first.

**Instruction to Patient:** "Mr. X, if we had 100 men exactly like you, we would expect between <predicted percentage from nomogram – 10%> and <predicted percentage + 10%> to remain free of their disease at 7 years following radical prostatectomy, and recurrence after 7 years is very rare."

A calibration plot for survival outcomes plots the Kaplan-Meier estimates at a certain time point against the predicted probabilities at the same time point. To calculate the Kaplan-Meier estimates, patients need to be grouped, for instance by deciles of predicted probability. Further, the slope of the calibration line can be estimated using a Cox regression analysis with recurrence as outcome variable and the linear predictor as only covariate.

Figure 2.6 shows the calibration plot of the Baylor nomogram, validated with the Cleveland data. Per decile of predicted probabilities, the Kaplan-Meier estimate and standard error were determined. The slope, calculated with Cox-regression, was 0.94 (95% confidence interval: 0.81 – 1.08).



**Figure 2.6** Validation plot of the Baylor nomogram with the Cleveland data. For each percentile of predicted probabilities, the average predicted probability is plotted against the Kaplan-Meier estimate. 95% confidence intervals of the Kaplan-Meier estimates are indicated with vertical lines (- -). The line approximates the ideal line (.....). Distribution of the predicted probabilities is indicated with vertical lines at the bottom (—). The probabilities lie on both sides of the threshold value (-.-).

Since a Cox model leaves the intercept free, the intercept of the calibration curve can not easily be calculated. To validate both the intercept and the slope, the time-axis needs to be transformed (for details see van Houwelingen *et al.*<sup>22</sup>). The resulting exponential model with linear predictor as only covariate yields an intercept and a slope.

To study ‘calibration-in-the-large’ for survival data, the Kaplan-Meier estimate of the complete sample at the time point of interest can be compared with the average predicted probabilities. In the Cleveland data, the Kaplan-Meier probability of being free of recurrence after 7 years was 73%. The average of all predicted probabilities, calculated with the nomogram, was 74%. Thus, observed frequencies and predicted probabilities were overall very similar.

‘Goodness-of-fit’ statistics like the H-L statistic for binary data are less straightforward for survival data.<sup>23</sup>

### Discrimination

As for calibration, the role of censoring should be carefully considered in the evaluation of the discriminative ability.<sup>24</sup> The distributions of the predicted probabilities by outcome (recurrence yes/no, censoring yes) are difficult to interpret, because the outcomes of the censored patients are not known. Drawing a ROC curve is problematic for the same reason. However, a *c*-statistic can still be calculated.<sup>25</sup> The interpretation of the *c*-statistic is similar to the interpretation for binary data. The *c*-statistic is the probability that from a random pair of patients the patient who recurred first had a higher predicted probability of recurrence.

The calculation assumes that the patient with the shorter follow-up time recurred first. If both patients recur at the same time or the non-recurrent patient has a shorter follow-up time, the pair of patients is discarded. The *c*-statistic was 0.83, when the Baylor nomogram was applied to the patients from Cleveland.

The distribution of the predicted probabilities can relatively easy be shown in the calibration plot (Figure 2.6) and gives some impression of the discriminative ability.

### *Clinical usefulness*

Censoring is also a complicating factor in the determination of measures related to clinical usefulness.<sup>24</sup> For each patient a predicted probability can be calculated, but in contrast to binary data the real outcome value is not known for censored patients. To establish a two by two table from which sensitivity and specificity can be calculated, the patients can still be divided in two groups using a threshold value for the predicted probabilities. The Kaplan-Meier curves of those groups show the observed frequencies of having the outcome at the considered time point. The number of patients having the outcome if all patients were followed for the total follow-up time can be calculated with these observed frequencies.

A sensible threshold value for the Baylor nomogram is not available yet, since there is no agreement on which additional treatment should be given and whether it would be effective. We therefore use a hypothetical threshold value of 30%. This threshold value implies that patients should receive early additional treatment, if the predicted probability of being free of recurrence after 7 years is lower than 30%. Furthermore, unnecessarily treating patients who will not relapse is 7/3 times worse than leaving patients untreated who will have progression of disease. The Kaplan-Meier estimates of being free of recurrence after 7 years were 6% for patients with a predicted probability lower than 30% and 80% for patients with a predicted probability above 30%. The two by two table corresponding to the 30% threshold value is shown in Table 2.4. Sensitivity, specificity and accuracy can now easily be calculated (Table 2.3).

To determine the clinical usefulness of the nomogram, the policy of not treating any patients ( $248 \times 1 = 248$  untreated while progression of disease [UPD]) was compared with treating patients with a predicted probability below 30% ( $173.2 \times 1 + 4.8 \times 7/3 = 184.4$  UPD). Thus, using the nomogram with a threshold value of 30% to select patients for early adjuvant treatment may be considered clinically useful (clinical usefulness =  $[248 - 184.4] / 248 = 26\%$ ) for the patients from the Cleveland Clinic in this study.

Further, the distribution of the predicted probabilities along with the indicated threshold value in the validation plot gives again an impression of the clinical usefulness (Figure 2.6).

**Table 2.4** Two by two tables for IUMC and Cleveland data to calculate sensitivity, specificity and accuracy*IUMC*

|                      | <i>Tumour<sup>a</sup></i> | <i>Benign tissue<sup>a</sup></i> | <i>Total</i> |
|----------------------|---------------------------|----------------------------------|--------------|
| Pred prob $\leq$ 70% | 193                       | 59                               | 252          |
| Pred prob $>$ 70%    | 7                         | 17                               | 24           |
| Total                | 200                       | 76                               | 276          |

<sup>a</sup> Outcome was observed at resection*Cleveland*

|                      | <i>Recurrence<sup>b</sup></i> | <i>No recurrence<sup>b</sup></i> | <i>Total</i> |
|----------------------|-------------------------------|----------------------------------|--------------|
| Pred prob $\leq$ 30% | 75.2                          | 4.8                              | 80           |
| Pred prob $>$ 30%    | 173.2                         | 692.8                            | 866          |
| Total                | 248.4                         | 697.6                            | 946          |

<sup>b</sup> Number of patients was estimated with Kaplan-Meier analysis**Discussion**

The performance of a prediction model can be assessed by studying three aspects of validity: calibration; discrimination; and clinical usefulness. Corresponding measures are applicable for internal and external validation studies, although the type of studies serve different purposes. Internal validity is mainly focussed on the development process and the quality of the model in a similar population. If the development sample was used for modelling decisions such as selection of variables and interaction terms, and non-linear relationships with the outcome, the model can perform excellently in the development data (apparent validity). However, calibration; discriminative ability; and clinical usefulness may be disappointing in resamples of the data (internal validity), because of the data-driven development process.<sup>26</sup> External validation studies are performed with new independent data to study the generalisability or transportability of the model. The external validity of a prediction model can be established by being cumulatively tested and found accurate across diverse settings, like any other scientific hypothesis.<sup>8, 27</sup> However, it is difficult to develop a model, which is well calibrated across diverse settings, since a model always consists of a limited number of predictors, leaving some variation between patients unexplained.

To study calibration the observed outcomes are compared with the expected probabilities, either with a calibration plot or the Hosmer-Lemeshow statistic (Table 2.1). Although the plot focuses on the calibration of a model, it can also give an impression of discriminative ability and clinical usefulness, when extended to a 'validation plot'. Therefore, a validation plot is the most informative tool described here. Particularly for survival data, the plot is preferable, because many measures have problems in dealing with



censoring. A disadvantage of the validation plot is that discrimination and clinical usefulness are not quantified. Comparing two models for those aspects will be difficult then.

The interpretation of the  $c$ -statistic depends on the clinical problem. The diagnosis of bacterial meningitis, for instance, can be established with a single test having a  $c$ -statistic above 0.80. To improve the diagnostic testing, a model which combines several test outcomes should achieve a  $c$ -statistic of at least 0.90.<sup>28</sup> On the other hand, accurate instruments for the selection of subfertile couples for IVF treatment are not available yet. A prediction model with a  $c$ -statistic of 0.65 may already be helpful for this clinical problem.<sup>29</sup>

The  $c$ -statistic is often used to indicate optimism. If the value decreases substantially in an independent data set, the model is considered as optimistic. However, the  $c$ -statistic also depends on the distribution of the predicted probabilities. A more homogeneous sample with patients having less extreme predictions will result in a lower  $c$ -statistic. This shows that the  $c$ -statistic should be carefully interpreted.

We briefly mention some overall performance measures, which estimate both calibration and discriminative ability. Decomposition of a performance measure into calibration and discrimination is best illustrated for the Brier score,<sup>9</sup> a measure for binary outcomes. The score estimates the mean difference between the observed outcome (0 or 1) and the predicted probability. For sensible models the Brier score ranges from 0 (perfect) to 0.25 (worthless). For the IUMC data the Brier score was 0.16. This result shows that the score is not easy to interpret. A related measure to the Brier score is  $R^2$ .  $R^2$  is interpretable as the proportion of outcome variation, which can be explained by the predictors of the model.

For the patients of IUMC  $R^2$  was 0.29. This value can be approximated by the difference in the average predicted probabilities of the two groups of patients with different outcomes.<sup>30</sup> For the patients from IUMC, the difference in average predicted probabilities was 0.25 (Figure 2.3A). For the Cleveland data,  $R^2$  was 0.10. These numbers suggests that the model for residual mass histology has a better performance than the Baylor nomogram, in contrast to other performance measures. This paradoxical result can partly be explained by the fact that for survival data the value of  $R^2$  depends on the number of censorings.<sup>31</sup>

Measures of clinical usefulness are rarely considered in model validation studies. This aspect has a different perspective than calibration and discrimination. It deals with the practical consequences of using a model, where the other two aspects focus on quantitative validity. In order to apply a model, a threshold value is necessary. Defining a threshold value is the most important step in assessing the clinical usefulness. Sometimes the relevant threshold value is clear, for instance if the aim of developing a prediction model is to predict death from severe head injury with 100% certainty.<sup>32</sup> In this case, all dying patients should have a predicted probability of dying of 100% and the threshold value is 100%. However, often a threshold value is less obvious. Disadvantages of classifying patients falsely as diseased or as non-diseased need to be balanced then. As we illustrated, the resulting threshold value can be used to calculate clinical usefulness as change in URM or UPD. We note that the considered selection of patients is important in the assessment of clinical

usefulness. In the testis example, we could only study patients who were selected for resection under the current policy. For those patients, the model was not useful. However, many patients were not selected for resection and for those the model might be valuable.

In conclusion, several aspects need to be considered in deciding whether a prediction model is useful for new patients. Calibration and discrimination are important to study whether the model provides reliable, and discriminating predictions. Furthermore, clinical usefulness indicates whether application of the model results in a better balance between patients falsely classified as diseased and falsely classified as non-diseased for the studied patient population.

## References

1. Harrell FE Jr, Lee KL, Mark DB: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 15: 361-387, 1996
2. Lemeshow S, Hosmer DW: Assessing the fit of the model, in *Applied logistic regression*. New York, Wiley, 1989, pp 135-175
3. Barndorff-Nielsen OE, Jensen JL (eds): *Statistical aspects of neural networks. Networks and Chaos - Statistical and probabilistic aspects*. London, Chapman and Hall, 1993
4. Chatfield C: Model uncertainty, data mining and statistical inference. *J R Statist Soc* 158:419-466, 1995
5. Efron B, Tibshirani R: Improvements on cross-validation: the .632+ bootstrap method. *JASA* 92:548-560, 1997
6. Steyerberg EW, Harrell FE Jr, Borsboom GJJ, et al: Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 54:774-781, 2001
7. Fletcher RH, Fletcher SW, Wagner EH (eds): *Clinical epidemiology. The essentials*. Baltimore, Williams & Wilkins, 1998
8. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. *Ann Intern Med* 130:515-524, 1999
9. Hilden J, Habbema JDF, Bjerregaard B: The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. *Meth Inform Med* 17:238-246, 1978
10. Steyerberg EW, Keizer HJ, Fosså SD, et al: Prediction of residual retroperitoneal mass histology following chemotherapy for metastatic nonseminomatous germ cell tumour: multivariate analysis of individual patient data from six study groups. *J Clin Oncol* 13:1177-1187, 1995
11. Vergouwe Y, Steyerberg EW, Foster RS, et al: Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. *J Urol* 165:84-88, 2001
12. Hastie T, Tibshirani R (eds): *Generalized additive models*. London, Chapman and Hall, 1990
13. Miller ME, Langefeld CD, Tierney WM, et al: Validation of probabilistic predictions. *Med Decis Making* 13:49-58, 1993
14. Cox DR: Two further applications of a model for binary regression. *Biometrika* 45:562-565, 1958
15. Spiegelhalter DJ: Probabilistic prediction in patient management and clinical trials. *Stat Med* 5:421-433, 1986
16. Hosmer DW, Hosmer T, le Cessie S, et al: A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 16:965-980, 1997
17. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143:29-36, 1982
18. Sox HC Jr, Blatt MA, Higgins MC, et al (eds): *Medical Decision Making*. Stoneham, Butterworth, 1988
19. Steyerberg EW, Marshall PB, Keizer HJ, et al: Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer. A decision analysis. *Cancer* 85:1331-1341, 1999
20. Habbema JDF: Clinical decision theory: the threshold concept. *Neth J Med* 47:302-307, 1995
21. Kattan MW, Wheeler TM, Scardino PT: Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *J Clin Oncol* 17:1499-1507, 1999
22. van Houwelingen HC, Thorogood J: Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med* 14:1999-2008, 1995
23. Parzen M, Lipsitz SR: A global goodness-of-fit statistic for Cox regression models. *Biometrics* 55:580-584, 1999

24. Graf E, Schmoor C, Sauerbrei W, et al: Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 18:2529-2545, 1999
25. Harrell FE Jr, Califf RM, Pryor DB, et al: Evaluating the yield of medical tests. *JAMA* 247:2543-2546, 1982
26. Steyerberg EW, Eijkemans MJC, Harrell FE Jr, et al: Prognostic modelling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 21:45-56, 2001
27. Altman DG, Royston P: What do we mean by validating a prognostic model? *Stat Med* 19:453-473, 2000
28. Spanos A, Harrell FE Jr, Durack DT: Differential diagnosis of acute meningitis. An analysis of the predictive value of initial observations. *JAMA* 262:2700-2707, 1989
29. Collins JA, Burrows EA, Willan AR: The prognosis for live birth among untreated infertile couples. *Fertil Steril* 64:22-28, 1995
30. Ash A, Schwartz M:  $R^2$ : a useful measure of model performance when predicting a dichotomous outcome. *Stat Med* 18:375-384, 1999
31. Schemper M, Stare J: Explained variation in survival analysis. *Stat Med* 15:1999-2012, 1996
32. Gibson RM, Stephenson GC: Aggressive management of severe closed head injury: time for reappraisal. *Lancet* ii:369-371, 1989

