Chapter 15
Evaluation of Performance

Background When we develop or validate a prediction model, we want to quantify how good the predictions from the model are ("model performance"). Predictions are absolute risks, which go beyond assessments of relative risks, such as regression coefficients, odds ratios, or hazard ratios. We can distinguish apparent, internally validated, and externally validated model performance (Chap. 5). For all types of validation, we need performance criteria in line with the research questions, and different perspectives can be chosen. We first take the perspective that we want to quantify how close our predictions are to the actual outcome. Next, more specific questions can be asked about calibration and discrimination properties of the model, which are especially relevant for prediction of binary outcomes in individual patients. We will illustrate the use of performance measures in the testicular cancer case study, with model development in 544 patients, internal validation with bootstrapping, and external validation with 273 patients from another centre.

15.1 Overall Performance Measures

The distance between the predicted outcome and actual outcome is a central to quantify overall model performance from a statistical perspective.[181] The distance is Y-$\hat{Y}$ for continuous outcomes. For binary outcomes, $\hat{Y}$ is equal to the predicted probability p, and for survival outcomes it is the predicted time to an event. These distances between observed and predicted outcomes are related to the concept of "goodness-of-fit" of a model, with better models having smaller distances between predicted and observed outcome.

15.1.1 Explained Variation: R^2

The amount of explained variation (R^2) is an overall measure to quantify the amount of information in a model in a given data set. R^2 is useful to guide various

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_15, © Springer Science+Business Media, LLC 2009

model development steps for all types of predictive regression models, including linear and generalized linear models (e.g. logistic, Cox). With R^2, we can readily compare the impact of different encoding of predictors, different shapes of the relationship of continuous predictors to the outcome, different selections of predictors, and the impact of including interaction terms (see previous chapters).

R^2 is the most common performance measure for continuous outcomes. For generalized linear models, Nagelkerke's R^2 can well be used.[309] As discussed in Chap. 4, this is a logarithmic scoring rule: $(Y - 1) - (\log(1 - p)) + Y \times \log(p)$. The logarithm of predictions p is compared with the actual outcome Y. For binary outcomes, the log likelihood for a patient with the outcome is $\log(p)$, without the outcome $\log(1 - p)$. When a very low prediction is made for a patient who actually had the outcome, this prediction has a severe score (Fig. 15.1). This may be a disadvantage for a prediction model that gives a prediction close to 0 or 1 while the outcome is discordant.

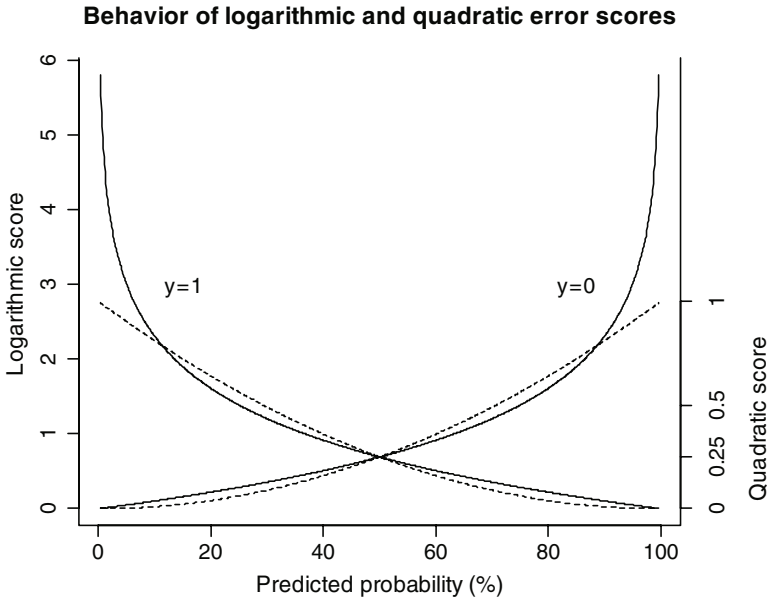

Fig. 15.1 Logarithmic and quadratic error scores of a subject with ($y = 1$) or without ($y = 0$) the outcome in relation to predicted probability (p). The logarithmic score was calculated as $y \times \log(p) + (1 - y) \times (1 - p)$, as in Nagelkerke's R^2 (solid line). The quadratic score was calculated as $(y - p)\hat{}2$, as in the Brier score (dashed line). Lines were scaled such that they crossed at $p = 50\%$. We note that the logarithmic score severely penalizes false predictions close to 0 or 100%

15.1.2 Brier Score

An alternative for binary outcomes is to use a quadratic scoring rule, where the squared differences between actual outcomes y and predictions p are calculated. This calculation is done in the Brier score, which is simply defined as $(Y - p)^2$. We can also write this similar as the logarithmic score: $Y \times (1 - p)^2 + (1 - Y) \times p^2$, with Y the outcome and p the prediction for each subject. For a subject, the score can range from 0 (prediction and outcome equal) to 1 (discordant prediction); a prediction of 50% has a score of 0.25 both when the outcome is 0 or 1. The Brier score is less severe than Nagelkerke's R^2 in penalizing false predictions close to 0% or 100% (Fig. 15.1). The Brier score for a model can range from 0% for a perfect model to 0.25 for a non-informative model with a 50% incidence of the outcome. When the incidence is lower, the maximum score for a model is lower, e.g. for 10%, 0.1 × (1 − 0.1)^2 + (1 − 0.1) × 0.1^2 = 0.090. A disadvantage of the Brier score is hence that the interpretation depends on the incidence of the outcome.

Similar to Nagelkerke's approach to the LR statistic, we could scale Brier by its maximum score: $\text{Brier}_{\text{scaled}} = 1 - \text{Brier} / \text{Brier}_{\text{max}}$, where $\text{Brier}_{\text{max}} = \text{mean}(p) \times (1 - \text{mean}(p))^2 + (1 - \text{mean}(p)) \times \text{mean}(p)^2$, with mean($p$) indicating the average probability of the outcome. $\text{Brier}_{\text{scaled}}$ ranges between 0% and 100%.

*15.1.3 Example: Performance of Testicular Cancer Prediction Model

We consider a development sample containing 544 patients contributed by six study groups,[417] and a validation sample 273 patients treated at Indiana University Medical Centre.[466] We developed a logistic regression model with five predictors: teratoma elements in the primary tumor, pre-chemotherapy levels of AFP and HCG, post-chemotherapy mass size, and reduction in mass size.

Internal validation of performance was estimated with bootstrapping (200 replications). Bootstrap samples were created by drawing random samples with replacement from the development sample. The prediction model was fitted in each bootstrap sample and tested on the original sample.

The essential *R* code is:

```
# 5 predictors in data set n544; develop model
full <- lrm(NEC ~ TER+PREAFP+PREHCG+SQPOST+REDUC10, data=n544)
val.prob(logit=full$linear.predictor, y=full$y) # apparent
validate(full, B=200)#Internal validation with 200 bootstraps
# External validation; refit model for matrix x and
# comparison of coefs
```

Table 15.1 Overall performance of testicular cancer prediction model

	Development	Internal validation	External validation
R^2	38.9%	37.6%	26.7%
Brier	0.174	0.178	0.161
$\text{Brier}_{\text{max}}$	0.248	0.248	0.201
$\text{Brier}_{\text{scaled}}$	29.8%	28.2%	20.0%

Development and internal validation with n=544 patients, external validation in n=273 patients. Internal validation with 200 bootstrap resamples using Harrell's validate function. $\text{Brier}_{\text{scaled}} = 1 - \text{Brier} / \text{Brier}_{\text{max}}$

```
ext.full <- lrm(NEC~TER+PREAFP+PREHCG+SQPOST+REDUC10,
            data=val, x=T, y=T)
lp <- ext.full$x %% full$coef [2:length(full$coef)] + full$coef[1]
val.prob(logit=lp,y=ext.full$y,riskdist="predicted")#external
```

Nagelkerke's R^2 was 38.9% in the development sample, and slightly lower at internal validation (Table 15.1). At external validation, the R^2 was estimated considerably lower, as 26.7%. Note that R^2 is based on the difference between a Null model ("intercept only") and a model with recalibrated predictions (intercept + calibration slope×logit of predictions).[174] So, the R^2 is estimated after recalibration of the predictions.

The Brier score was 0.174 and 0.178 at development and internal validation respectively. Remarkably, the Brier score was better at external validation (0.161). The external Brier score was simply calculated by comparing predictions with actual outcome, without recalibration as was done for R^2. The interpretation of the Brier score is easier with the scaled version, which compensates for the fact that the maximum Brier score was lower in the external validation set (necrosis in 76 of 273 (28%); $\text{Brier}_{\text{max}}$, 0.20) than in the development set (necrosis in 245 of 544 (45%); $\text{Brier}_{\text{max}}$, 0.25). The scaled Brier score was clearly lower at external validation than at internal validation (20% vs. 28%, Table 15.1).

*15.1.4 Overall Performance Measures in Survival

Nagelkerke's R^2 can readily be calculated for survival outcomes, based on the difference in −2 log likelihood of a model without and a model with the linear predictor. Calculation of the Brier score is not directly possible because of censoring: Not all subjects are followed long enough for the outcome to occur. To address the censoring issue, we can define a weight function, which considers the conditional probability of being uncensored during time.[146,375,374] The assumption is that the censoring mechanism is independent of survival and the subject's history.

Table 15.2 Classification of subjects according to a cutoff for the probability of an outcome (event or no event)

	Event	No event
Predicted probability >= cutoff	TP	FP
Predicted probability < cutoff	FN	TN
	N_{event}	$N_{\text{no event}}$

TP and FP: Numbers of true and false-positive classifications; FN and TN: Numbers of false and true-negative classifications, respectively. N_{event} = TP + FN; $N_{\text{no event}}$ = FP + TN

We can hence calculate the Brier score at fixed time points. For example, we can compare predicted survival vs. observed survival at 1, 2, and 5 years of follow-up. Choosing many consecutive time-points leads to a time-dependent graph. This is useful to use a benchmark curve, based on the Brier score for the overall Kaplan-Meier estimator, which does not consider any predictive information. The survival estimates of the overall Kaplan-Meier curve only depend on time of follow-up, and are identical for all subjects alive at a certain point in time. An interesting example is provided by a case study on the disappointing contribution of microarray data to prediction of survival for patients with diffuse large-B-cell lymphoma.[374]

*15.1.5 Decomposition in Discrimination and Calibration

Overall statistical performance measures incorporate both calibration and discrimination aspects. For example, the Brier score can formally be decomposed into indicators of calibration and discrimination.[303,38] Discrimination relates to how well a prediction model can discriminate those with the outcome from those without the outcome. Calibration relates to the agreement between observed outcomes and predictions. Studying discriminative ability and calibration is often more meaningful than an overall measure such as R^2 or Brier score when we want to appreciate the quality of model predictions for individuals. We therefore discuss these aspects further.

15.1.6 Summary Points

- R^2 is a common measure to express the amount of variability in outcomes that is explained by the prediction model
- The Brier score is another common performance measure for the distance between observed and predicted outcome, which can be decomposed in discrimination and calibration aspects

15.2 Discriminative Ability

Model predictions need to discriminate between those with and those without the outcome (Event vs. No event). Several measures can be used to indicate how good we classify patients in a binary prediction problem. The concordance (*c*) statistic is the most commonly used performance measure to indicate the discriminative ability of generalized linear regression models. For a binary outcome *c* is identical to the area under the receiver operating characteristic (ROC) curve. The ROC curve is a plot of the sensitivity (true positive rate) against 1 – specificity (false-positive rate) for consecutive cutoffs for the probability of an outcome. We therefore consider sensitivity and specificity first.

15.2.1 Sensitivity and Specificity of Prediction Models

Sensitivity is defined as the fraction of true-positive (TP) classifications among the total number of patients with the outcome (TP/N_{event}), and the specificity as the fraction of true-negative classifications among the total number of patients without the outcome ($TN/N_{no\ event}$, Table 15.2). To classify a patient as positive or negative, we need to apply a cutoff to the predicted probability. If the prediction is higher than the cutoff, the patient is classified as positive, otherwise as negative. It is common to use a cutoff of 50% for classification. This cutoff is often not defendable in a medical context, as we will discuss in detail in the next chapter (Chap. 16). We can examine sensitivity and specificity over the whole range of cutoffs from 0% to 100%. The results can be plotted in an ROC curve.[172]

15.2.2 Example: Sensitivity and Specificity of Testicular Cancer Prediction Model

If we classify patients as having necrosis when the probability of necrosis is over 50%, we have a sensitivity of 68% and a specificity of 77% (FP rate, 23%). With a higher cut-off, for example 70%, these numbers are 42% and 92%, respectively. This illustrates that a higher cutoff leads to better specificity, at the price of a lower sensitivity. This trade-off is visualized in an ROC curve (Fig. 15.2).

15.2.3 ROC Curve

A plot of an ROC curve has often been used in diagnostic research to quantify the diagnostic value of a test over its whole range of possible cutoffs for classifying patients as positive vs. negative. We can also make an ROC curve with consecutive cutoffs for the predicted probability of a binary outcome. We start with a cutoff of

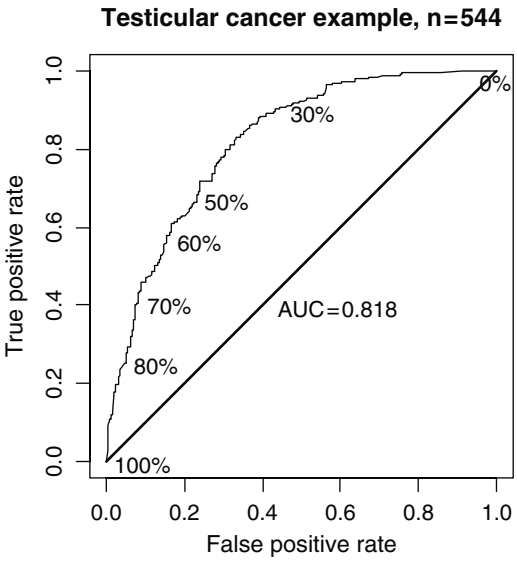

Fig. 15.2 Receiver operating characteristic (ROC) curve for the testicular cancer model in the development data set of 544 patients. Using cutoffs for the predicted probability of necrosis (benign tissue) results in specific combinations of true-positive rate (sensitivity) and false-positive rate (1 – specificity). The area under the curve is 0.818

0%, which implies that all subjects are classified as positive. The sensitivity is 100%, and the specificity 0% (upper-right point in Fig. 15.2). There are no false-negative classifications, and 100% false-positive classifications, since all subjects without the outcome are classified as positive. We then shift to a slightly higher cutoff, e.g. 1%, where sensitivity may still be 100%, but specificity above 0%. We follow all possible cutoffs till 100%, where all subjects are classified as negative. This is the lower-left point in Fig. 15.2. The sensitivity is then 0%, and specificity 100%. The curves are more to the upper left corner when the distributions of predictions are more separate between those with and without the outcome (Fig. 15.3).

We can draw a line between the 0%, 0% and 100%, 100% points, indicating a non-informative model. Note that the sum of TP and TN is 1 at every cutoff for such a model. This sum (also known as Youden's index) is larger than 1 for sensible prediction models.

The area under the curve can be interpreted as the probability that a patient with the outcome is given a higher probability of the outcome by the model than a randomly chosen patient without the outcome.[172] An uninformative model, such as a coin flip, will hence have an area of 0.5. A perfect model has an area of 1. The interpretation hence is relatively straightforward, but assumes that we have a pair of patients, one with and one without the outcome. This is a rather artificial situation. Statistically, this conditioning on a pair of patients is attractive, since it makes the area independent of the incidence of the outcome, in contrast to R^2 or the Brier score for example.

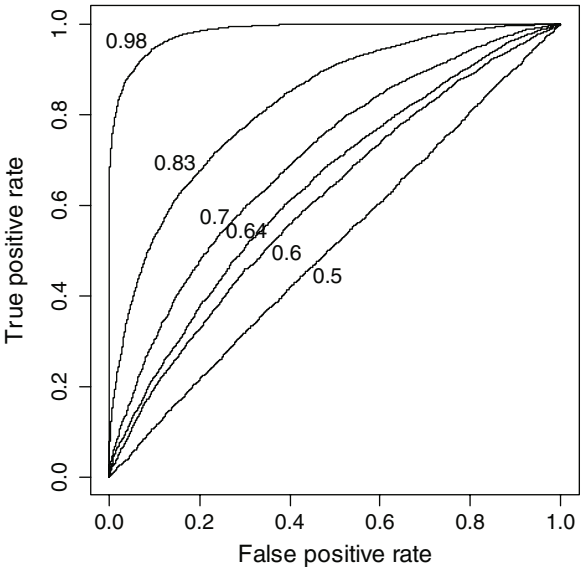

Fig. 15.3 ROC plot for five hypothetical prediction models. Models were created with distributions as shown in Fig. 15.4 (see also Fig. 4.6). The *c* statistics were 0.5, 0.6, 0.64, 0.7, 0.83, and 0.98 at 50% incidence of the outcome

A generalization of the area under the ROC curve is provided by the concordance statistic (*c*).[175] The *c* statistic is a rank order statistic for predictions against true outcomes, related to Somer's D statistic. As a rank order statistic, it is insensitive to errors in calibration such as differences in average outcome. For binary outcomes, *c* is identical to the area under the ROC curve.

Confidence intervals for the area under ROC curve (or *c* statistic) can be calculated with various methods. Standard asymptotic methods may be problematic, especially when sensitivity or specificity are close to 0% or 100%.[9] Bootstrap resampling is a good choice for many situations. For example, differences in *c* between models fitted on the same data can be tested with standard formulas for the difference. But such formulas are only valid if the models were pre-specified. If one or both models were estimated on the same data, bootstrapping can be used for comparison of optimism-corrected estimates (see Chap. 17).

15.2.4 R^2 vs. c

We compare the behavior of Nagelkerke's R^2 and the *c* statistic in some simulations over a range of incidences of the outcome (1%, 10%, 50%, 90%, Fig. 15.4). At 50% incidence, a high *c* statistic such as 0.98 is associated with an R^2 value of 87%. With lower incidence, R^2 is somewhat lower.

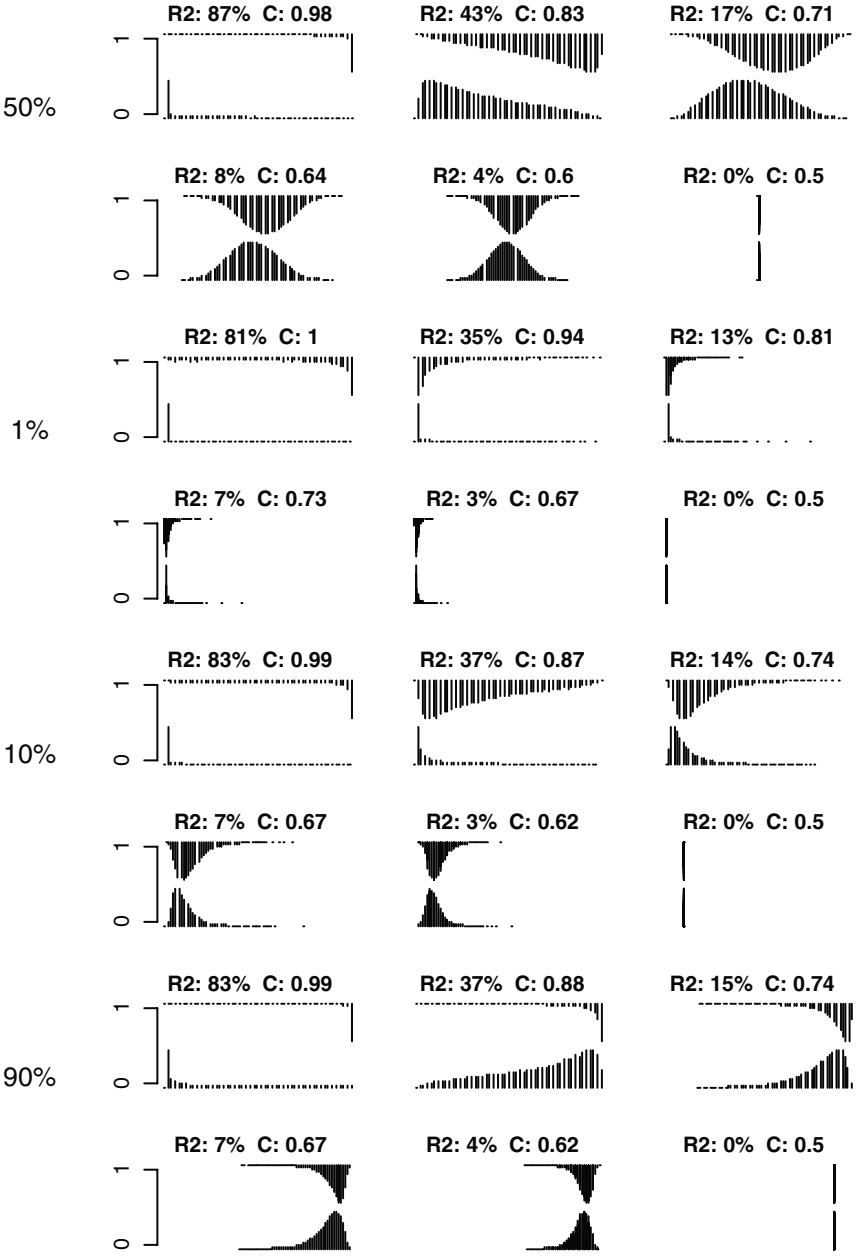

Fig. 15.4 Distribution of observed outcomes (0 or 1), in relation to predicted probabilities from hypothetical logistic models relating Y to a predictor X. The *top* six graphs relate to an incidence of 50%. The next sets of 3 × 6 graphs relate to incidences of 1%, 10%, and 90% respectively. For each hypothetical model, Nagelkerke's R^2 and c statistic are listed. If c=0.5 (and R^2=0%), predictions are at the incidence of the outcome for all subjects, with or without the outcome, indicated with a single spike. If c is close to 1 (R^2 close to 100%), predictions are close to 0% for those without the outcome, and close to 100% for those with the outcome. Note that R^2 and c statistics differ somewhat between 10% and 90% incidence, because of random noise in the simulation procedure

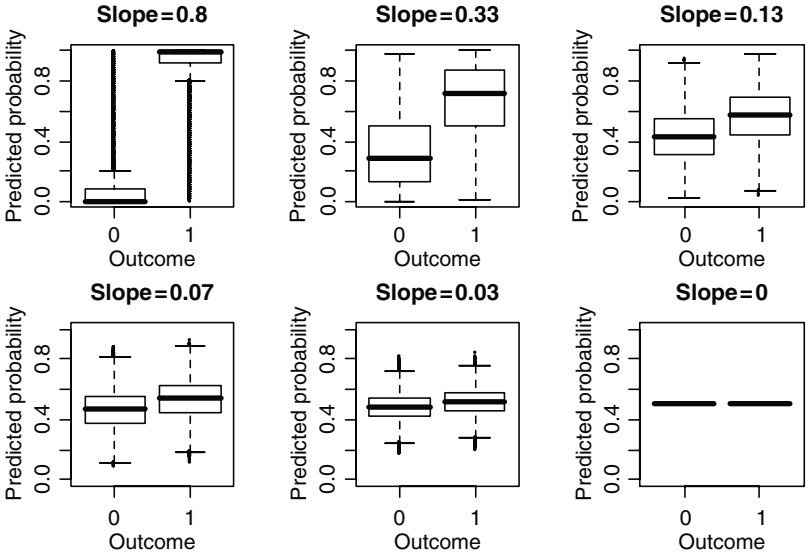

Fig. 15.5 Box plots for predictions from six hypothetical prediction models with different discriminative ability (see Fig 15.4). The discrimination slopes are calculated as the difference in means of predictions for those with and those without the outcome (mean incidence, 50%)

*15.2.5 Box Plots and Discrimination Slope

The discrimination slope has been proposed as a simple measure for how well subjects with and without the outcome are separated. It is easily calculated as the absolute difference in average predictions for those with and without the outcome.

Visualization is readily possible with a box plot (Figs. 15.5 and 15.7). The box plot may be a simple and intuitive way to communicate the extent of risk differentiation achieved by the model. The same information can be shown by histograms, which will show less overlap between those with and those without the outcome for a better discriminating model (Fig. 15.4). Similar to Fig. 15.4, the incidence of the outcome determines the visual expression that a box plot makes, and the magnitude of the discrimination slope. With low incidence, the slope is somewhat lower, for the same *c* statistic.

*15.2.6 Lorenz Curve

An alternative way to judge discriminative ability is the Lorenz curve (Fig. 15.6). The Lorenz curve has been used in economics to characterize the distribution of wealth in a population.[267] This curve has been used to plot the cumulative distribution of wealth against the cumulative distribution of the population, ranked on the basis of individual wealth.

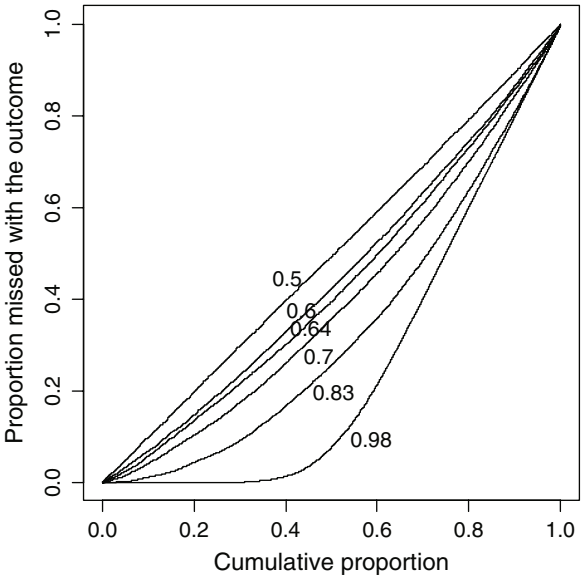

Fig. 15.6 Lorenz curve showing proportion missed with the outcome vs. the cumulative proportion of patients according to rank order of predictions, for an outcome incidence of 50%. We note that a near perfect model (c=0.98) follows a horizontal line and then rises steeply to 100% false-negative rate from the point of 50% cumulative proportion.

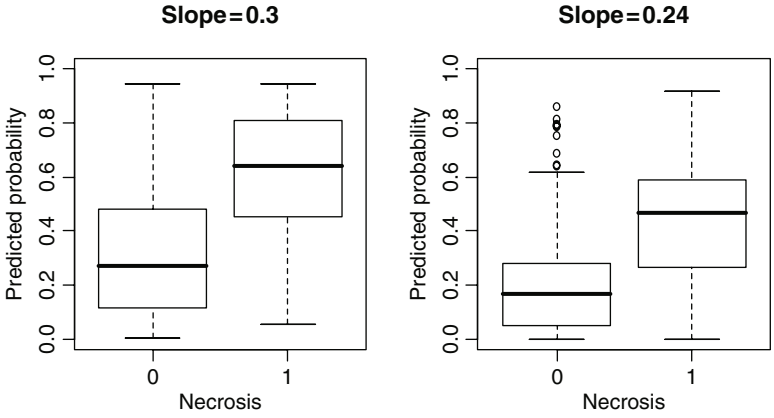

Fig. 15.7 Box plot showing predictions by actual outcome (necrosis) for testicular cancer patients (n=544 and 273, respectively)

Table 15.3 Summary of some measures for discriminative ability of a prediction model for binary outcomes

Measure	Calculation	Visualization	Pros	Cons
Concordance statistic	Rank order statistic	ROC curve	Insensitive to outcome incidence; interpretable for pairs of patients with and without the outcome	Interpretation artificial
Discrimination slope	Difference in mean of predictions between outcomes	Box plot	Easy interpretation, nice visualization	Depends on the incidence of the outcome
Lorenz curve	Shows concentration of outcomes missed by cumulative proportion of negative classifications	Concentration curve	Shows balance between finding true positive subjects vs. total classified as positive	Depends on the incidence of the outcome

For prediction models we can plot the cumulative proportion of the population on the x axis, ranked by predicted probability. On the y axis, we plot the cumulative proportion of subjects with the outcome. For example, we can show the proportion of subjects developing cancer against the cumulative proportion of the population ranked by cancer risk.[31] In terms of ROC curves, we plot the cumulative rate of false-negative classifications against the total of negative predictions. With incidences of the outcome around 50%, the ROC and Lorenz curves look very similar, except that the Lorenz curve is flipped vertically and horizontally. In case of a non-informative model, a straight line arises, since every rate of the population classified as negative corresponds to the same rate classified as negative among those with the outcome. A good model has a curve under this straight line, with a relatively large proportion of the population classified as negative having only a small part of the outcomes (low false-negative rate). On the upper end of the x axis, a small part of the population should contain many subjects with the outcome. In the ideal case, a cutoff is used that classifies the fraction as positive, equal to the prevalence, and all these have the outcome. Indeed, we note that a c statistic of 0.98 leads to a nearly horizontal line till the 50% cumulative proportion point on the x axis, and increases more or less linearly to 100% after that.

The Gini index is often calculated as a summary measure for the Lorenz curve. The Gini index is the ratio between the area (A) between the Lorenz curve of the prediction model and the line for a non-informative model and the area under the line for an non-informative model (0.5). Hence, $G = 2A$.

Other summaries are related to quantiles of the cumulative distribution. For example, we can consider the number of missed outcomes when 25% of the population is classified as negative. If we want to be sure not to miss the outcome, usually only

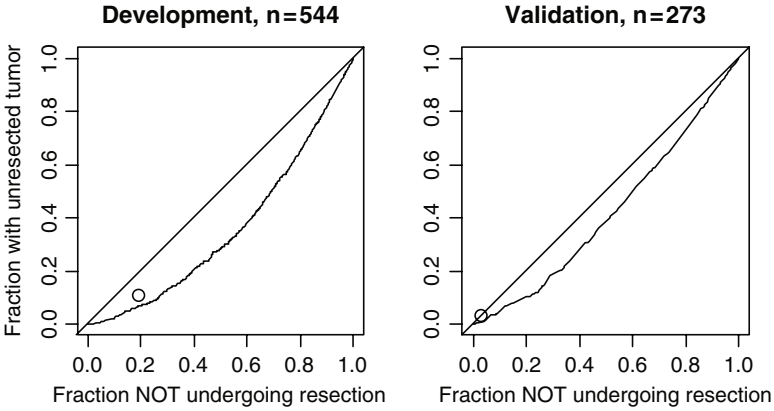

Fig. 15.8 Lorenz curves for prediction of necrosis vs. residual tumor. Patients classified as necrosis would not undergo surgical resection (*x* axis). With increasing fractions not undergoing resection, the fraction with unresected tumor increases ("missed tumor"). With 75% undergoing resection, 56% of the tumors are resected, leaving 44% unresected

few can be classified as negative, unless a model is used with very good discriminative ability. At the upper end of the range, we can consider how many outcomes are concentrated in the upper quartile (above 75 percentile). We will illustrate these percentiles for the testicular cancer prediction case study (Fig. 15.8).

An advantage of the Lorenz concentration curve is that the trade-off is clearly visualized between how many subjects can be classified as negative without missing many with the outcome. A disadvantage is that the appearance of the Lorenz curve depends strongly on the incidence of the outcome; with low incidence, the graph looks impressive, and with high incidence, the graph looks rather poor. As an example, consider a screening setting with 1% of subjects having the disease of interest. Only few cases with disease are missed at 25% classified negative when we use a model with a *c* statistic of 0.83. The top 25% then easily contains most cases. With a more frequent outcome, more cases are missed at the point of 25% classified negative, and fewer of the cases are in the top 75 percentile.

15.2.7 Discrimination in Survival Data

For survival data, Harrell's overall *c* statistic indicates the proportion of all pairs of subjects who can be ordered such that the subject with the higher predicted survival is the one who survived longer.[175] Ordering is possible if both subjects have an observed survival time, or when one has the outcome and a shorter survival time than the censored survival time of the other subject. Ordering is not possible if both

subjects are censored, or if one has the outcome with a survival time longer than the censored survival time of the other subject. Some alternative definitions of *c* have been proposed, which lead to time-dependent performance curves.[183]

In oncology, prognostic groups are often created after constructing a prognostic model. A common procedure is to base these groups on quartiles of predicted survival; the lower 25% should have the worst survival and the highest 25% the best survival. This approach can well illustrate the discriminative ability of a model. An example is shown in Chap. 23 (Fig. 23.8).

15.2.8 Example: Discrimination of Testicular Cancer Prediction Model

We continue the example of predicting a benign histology in testicular cancer patients after chemotherapy. The *c* statistic was 0.818 at model development, with small optimism according to bootstrap validation (decrease by 0.006 to 0.812). At external validation, the *c* statistic was 0.785, with a relatively wide 95% confidence interval of 0.73 to 0.84 (Table 15.4).

The discrimination slope was 0.30 at model development, with small optimism according to bootstrap validation (decrease to 0.29). At external validation, the slope was much smaller (0.24). Part of this decrease is attributable to the lower average prevalence of necrosis (76 of 273, 28%, vs. 245 of 544, 45%). This lower prevalence is also evident from the box plots (Fig. 15.7).

The Lorenz curves were created with *x* axis as the cumulative fraction classified as necrosis, i.e. not having tumor, and hence classified as not undergoing surgical resection (Fig. 15.8). The *y* axis was the fraction of missed tumors, i.e. tumor masses left unresected. The point of 25% classified as necrosis corresponds to using a cutoff of 68% for the probability of necrosis; only patients with

Table 15.4 Discriminative ability of testicular cancer prediction model

	Development ($n=544$, 245 necrosis)	Internal validation	External validation ($n=273$, 76 necrosis)
c statistic [95% CI]	0.818 [0.783–0.852]	0.812 [0.777–0.847][a]	0.785 [0.726–0.844]
Discrimination slope [95% CI]	0.301 [0.235–0.367][b]	0.294 [0.228–0.360][a]	0.237 [0.178–0.296][b]
Lorenz curve p25, tumors missed	9%	–	13%
Lorenz curve p75, tumors missed	58%	–	65%

Development and internal validation with $n=544$ patients, external validation in $n=273$ patients. Internal validation with 200 bootstrap resamples using Harrell's `validate` function

[a]Assuming the same SE applies as estimated for model development

[b]Based on bootstrap resampling

a probability over 68% are not resected. We miss 9% of the tumors with that cutoff. Hence, sparing surgery in 25% leads to missing 9% of the tumors. The point of 75% classified as necrosis corresponds to using a low cutoff (21%), and missing 58% of the tumors. Hence 42% of the tumors are concentrated in the upper quartile of the distribution.

At external validation, the curve looks worse, which is related to a lower discriminative ability and to a lower average prevalence of necrosis (28% vs. 45%). The 25% and 75% cumulative fractions correspond to cutoffs of 40% and 8% for the probability of necrosis, and lead to 13% and 65% missed tumors, respectively.

As a reference, we consider the current widely used policy of resection if the residual mass size exceeds 10 mm.[418] This policy uses only one of the five predictors in the model (post-chemotherapy mass size), and hence has less discriminative ability (the point is closer to the 45° line in Fig. 15.8). In the development sample, 107 of the 544 patients (20%) had residual masses <= 10 mm, but among them 30 with tumor (fraction tumor missed, 30 of 299, 10%). In the validation sample, only 9 of the 273 patients (3.3%) had residual masses <= 10 mm, but among them, 6 with tumor (fraction tumor missed, 6 of 197, 3%). Hence, the reference policy did not perform well in the validation sample.

*15.2.9 Verification Bias and Discriminative Ability

In the testicular cancer validation sample, only nine patients had very small residual masses. This reflects the policy for resection in the specific centre, where patients with such very small masses were not considered candidates for resection.[466] This leads to verification bias; we do not know the histology of these masses, since they were not resected, and cannot evaluate predictions for these patients. We know that the estimation of regression coefficients is not biased by this selection, if we include the selection criterion (residual mass size) in the prediction model. Hence model predictions are valid even with verification bias.[497] But performance measures such as sensitivity and specificity suffer from this verification bias.[30] The *c* statistic may not be affected too much because verification bias makes that we merely shift on the ROC curve to a different combination of sensitivity and specificity.

*15.2.10 R Code

The boxplot is created simply with the `boxplot` command, based on a "full model," including five predictors in the development data:

```
lp <- full$linear.predictors
boxplot(plogis(lp ~ full$y)          # Fig 15.7
```

The discrimination slope is the difference between the mean predicted probabilities by outcome:

```
mean(plogis(lp[full$y==1])) - mean(plogis(lp[full$y==0]))
```

Lorenz curves are created with the ROCR package:

```
library(ROCR)
# Make ROC object with predicted probability for outcome
pred.full <- prediction(plogis(lp), full$y)
# Lorenz curve data and plot
perf1      <- performance(pred.full, "fpr", "rpp")
plot(perf1, xlab="NOT undergoing resection",
            ylab="with unresected tumor")
abline(a=0, b=1)       # Fig 15.8
```

15.3 Calibration

Another important property of a prediction model is calibration, i.e. the agreement between observed outcomes and predictions. For example, if we predict 70% probability of benign tissue for a testicular cancer patient, the observed frequency of benign tissue should be 70 out of 100 such patients.

15.3.1 Calibration Plot

A calibration plot has predictions on the *x* axis, and the outcome on the *y* axis. A line of identity helps for orientation: Perfect predictions should be on the 45° line. For linear regression, the calibration plot results in a simple scatter plot. For binary outcomes, the plot contains only 0 and 1 values for the *y* axis. Probabilities are not observed directly. However, smoothing techniques can be used to estimate the observed probabilities of the outcome ($p(y=1)$) in relation to the predicted probabilities. The observed 0/1 outcomes are replaced by values between 0 and 1 by combining outcome values of subjects with similar predicted probabilities, e.g. using the loess algorithm.[174] We can also plot results for subjects grouped by similar probabilities (quantiles), and thus compare the mean predicted probability to the mean observed outcome. For example, we can plot observed outcome by decile of predictions (Fig. 15.9). This makes the plot a graphical illustration of the Hosmer-Lemeshow goodness-of-fit test (see Sect. 15.3.8 and 15.3.10). A better discriminating model has more spread between such deciles than a poorly discriminating model. The choice of quantiles is important for the visual impression of calibration; if small groups are plotted, the variability will be large.

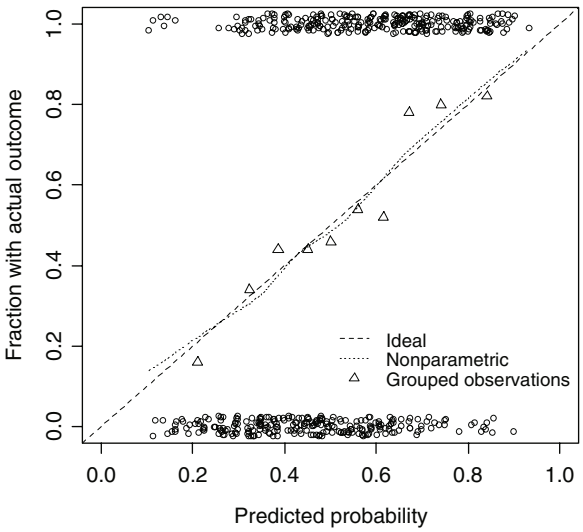

Fig. 15.9 Calibration plot of actual outcome vs. predictions for a hypothetical model with *c* statistic 0.7, *n*=500. The distributions of actual 0 and 1 values are shown at the *bottom* and at the *top* of the graph; the loess smoother is close to the ideal 45° line; actual outcomes by deciles of risk are shown by triangles (each triangle, $n=50$)

15.3.2 Calibration in Survival

In a survival context, the calibration of a model is usually studied at fixed time points. For these time points, we can consider grouped patients, with sufficient numbers per group to allow for calculation of survival rates with the Kaplan-Meier method. This observed survival is compared with the mean predicted survival from the prognostic model. Harrell suggests to use at least 50 subjects per group, depending on the hazard of the outcome.[174] It would be interesting to plot a smoothed curve as for binary outcomes, but this is not easy.

15.3.3 Calibration-in-the-Large

A calibration plot can easily be made for the data set used to develop a model. This indicates the apparent calibration. In model development, the average of predictions is the average of the outcomes: mean (Y) = mean($\hat{Y}$). For example, mean(observed BP) = mean(predicted BP) in linear regression, and mean(observed 30-day mortality) is mean(predicted 30-day mortality). This correspondence is guaranteed by the intercept in a (generalized) linear model. This correspondence of average outcomes remains at internal validation with bootstrapping. When we apply the model to external data, this correspondence may be less. The difference between mean($\hat{Y}$) and mean(Y_{new}) is referred to as "calibration-in-the-large."

15.3.4 Calibration Slope

Another important calibration measure is related to the average strength of the predictor effects. For linear regression, we can write $Y_{new} = a + b_{overall}\ \hat{Y}$, and for generalized linear models $f(Y_{new}) = a + b_{overall}$ linear predictor, where the linear predictor is the combination of regression coefficients from the model and the predictor values in the new data. A link function f is used for Y_{new}, e.g. logodds (or logit) in logistic regression. The $b_{overall}$ is named the calibration slope.[86] Ideally, the calibration slope $b_{overall} = 1$. With apparent validation, $b_{overall} = 1$ because this yields the best fit on the data under study with either least squares or maximum likelihood methods. At internal validation, the calibration slope reflects the amount of shrinkage that is required for a model ($b_{overall} < 1$).[81] It indicates how much we need to reduce the effects of predictors on average to make the model well calibrated for new patients from the underlying population. The calibration slope can hence be used as a shrinkage factor to adjust a model for future use (Chap. 14). At external validation, the calibration slope reflects the combined effect of two issues: overfitting on the development data and true differences in effects of predictors.

15.3.5 Estimation of Calibration-in-the-Large and Calibration Slope

For continuous outcomes, calibration-in-the-large can be assessed easily by comparing the mean ($\hat{Y}$) and mean(Y_{new}), and testing the differences $Y_{new} - \hat{Y}$, e.g. with a one-sample t-test. This test indicates the statistical significance of the mean under- or overestimation of the observed outcome: mean($Y_{new} - \hat{Y}$). In a linear regression model, we can estimate an intercept a in the model with as outcome the residual $Y_{new} - \hat{Y} : Y_{new} - \hat{Y} =$ a. The recalibration model is simply $Y_{new} = a + b_{overall}\ \hat{Y}$. The deviation of the calibration slope from 1 can be tested in linear regression by a model that studies the residuals: $Y_{new} - \hat{Y} = a + b_{overall}\ \hat{Y}$. The significance of $b_{overall}$ is then determined as usual in regression, and indicates on average stronger or weaker effects of the predictors in a model.

For binary outcomes, calibration-in-the-large again refers to the difference between mean $\hat{Y}$ and mean(Y_{new}). A simple comparison can directly be made, with an odds ratio indicating the average under- or overestimation of the outcome:

$$\text{OR} = \text{odds}(\text{mean}(\hat{Y}))/\text{odds}(\text{mean}(Y_{new})) = [\text{mean}(\hat{Y})/(1-\text{mean}(\hat{Y})]/[\text{mean}(Y_{new})/(1-\text{mean}(Y_{new})].$$

For statistical testing of the difference we need to be more careful. In logistic regression, the relationship between the outcome y and the linear predictor is non-linear (i.e. logistic). We have to compare logit($Y_{new} = 1$) to logit($\hat{Y}$), where mean (logit($Y_{new} = 1$) – logit ($\hat{Y}$)) is not equal to mean(logit($Y_{new} = 1$)) – mean (logit ($\hat{Y}$)).

In a model, we could write

$$\text{logit}(Y_{new} = 1) - \text{logit}\ (\hat{Y}) = \text{a};$$

$$\text{or logit}(Y_{new} = 1) = a + \text{logit}\,(\hat{Y}) = a + \text{offset (linear predictor)}.$$

The intercept a then reflects the difference in logodds between predictions and observed outcome, adjusted for the linear predictor. The offset makes that predictions are taken literally, as in linear regression. Values of the offset variable are subtracted from the actual outcomes Y_{new} (as in Poisson regression). Equivalently we can think of a regression coefficient for the offset variable that is fixed at unity. The statistical significance of intercept a can be tested with standard regression tests, such as the Wald test or the likelihood ratio (LR) test.

The calibration slope can be estimated from the recalibration model

$$\text{logit}(Y_{new} = 1) = a + b_{overall} \times \text{logit}\,(\hat{Y}) = a + b_{overall} \times \text{linear predictor}.$$

The deviation of the calibration slope from 1 ("miscalibration") can be tested by a model that includes an offset variable:

$$\text{logit}(Y_{new} = 1) = a + b_{miscalibration} \times \text{linear predictor} + \text{offset (linear predictor)}.$$

The slope coefficient $b_{miscalibration}$ reflects the deviations from the ideal slope of 1, and can be tested with Wald or LR statistics.

Calibration-in-the-large cannot be detected with a refitted Cox regression model, since the baseline hazard h_0 is usually left free in fitting such a model. For a survival outcome, the calibration slope can be assessed as:

$$\text{log(hazard}(y_{new} = 1)) = h_0 + b_{overall} \times \text{linear predictor}.$$

The model for deviation from a slope of 1 is:

$$\text{log(hazard}\,(y_{new} = 1)) = h_0 + b_{miscalibration} \times \text{linear predictor} + \text{offset (linear predictor)}.$$

Testing of coefficient $b_{miscalibration}$ is as usual, i.e. with a Wald test or LR test.

With a parametric survival model, we can specify parameters that reflect differences in average survival, after adjustment for predictor effects. Van Houwelingen hereto transformed the baseline hazard from a Cox model to a Weibull model.[456] The Weibull model has two parameters to describe the baseline hazard parametrically (Chap. 4). These two parameters can be refitted for external validation data, together with the linear predictor, to estimate a recalibrated model.

*15.3.6 Other Calibration Measures

Various other measures are available for calibration. An intuitively appealing measure of calibration is the absolute difference between smoothed observed outcomes

Table 15.5 Calibration tests for prediction model $y \sim a + b_{overall}\,\hat{y}$

	H_0	H_1	*df*
Calibration-in-the-large	$a=0 \mid b_{overall} = 1$	$a<>0 \mid b_{overall} = 1$	1
Calibration slope	$b_{overall} = 1$	$b_{overall} <> 1$	1
Recalibration	$a = 0$ and $b_{overall} = 1$	$a <> 0$ or $b_{overall} <> 1$	2

H_0 and H_1 indicate the Null and alternative hypothesis respectively

and predicted probabilities (Harrell's *E* statistic).[174] This measure is related to the calibration plot, and depends on the way the 0/1 outcomes are smoothed. The difference between smoothed observed outcomes and predicted probabilities can also be judged visually in a calibration plot such as Fig. 15.9.

15.3.7 Calibration Tests

Statistical tests can be performed with various null hypotheses for calibration, phrased in the formulation of the recalibration model $y \sim a + b_{overall}\,\hat{Y}$ (Table 15.5). Tests for calibration-in-the-large and calibration slope have one *df*; the calibration test has two *df*. The test for calibration-in-the-large requires that the predictions are taken literally ($b_{overall} = 1$). In generalized linear models, this can be achieved with an offset variable. The calibration slope can easily be estimated in the recalibration model. The recalibration test has several advantages (Table 15.6). It can pick-up common patterns of miscalibration, i.e. systematic differences between the new data and the model development data, and overfitting of the effects of predictors. Moreover the test parameters a and $b_{overall}$ are well interpretable, provided that $a \mid b_{overall} = 1$ is reported (rather than a with $b_{overall}$ left free). The slope $b_{overall}$ can directly be taken from the re-calibration model (where a is left free).

Statistical testing for calibration has a number of drawbacks. First, the null hypothesis is of good calibration. Hence, if we test calibration in a small study, we have low power and will not reject the null hypothesis unless miscalibration is very severe. On the other hand, even a model with very good, but not perfect, calibration will fail if the sample size is sufficiently large.

15.3.8 Goodness-of-Fit Tests

Calibration is related to goodness-of-fit, which relates to the ability of a model to fit a given set of data. Typically, there is no single goodness-of-fit test that has good power against all kinds of lack of fit of a prediction model. Examples of lack of fit are missed non-linearities, interactions, or an inappropriate link function between the linear predictor and the outcome. Goodness-of-fit can be tested with a χ^2 statistic.

For binary outcomes, the Hosmer-Lemeshow (H-L) goodness-of-fit test is often used.[199] Usually, patients are grouped by decile of predicted probability. The sum

Table 15.6 Summary of some measures for calibration of a prediction model for binary outcomes

Performance aspect	Calculation	Visualization	Pros	Cons
Calibration-in-the-large	Compare mean(y) vs. mean($\hat{y}$)	Calibration graph	Key issue in validation; statistical testing possible	By definition OK in model development setting
Calibration slope	Regression slope of linear predictor	Calibration graph	Key issue in validation; statistical testing possible	By definition OK in model development setting
Calibration test	Joint test of calibration-in-the-large and calibration slope	Calibration graph	Efficient test of two key issues in calibration	Insensitive to more subtle miscalibration
Harrell's E Statistic	Absolute difference between smoothed y vs. line of identity	Calibration graph	Conceptually easy, summarizes miscalibration over whole curve	Depends on smoothing algorithm
Hosmer-Lemeshow test	Compare observed vs. predicted in grouped patients	Calibration graph or table	Conceptually easy	Interpretation difficult; low power in small samples
Goeman–Le Cessie test	Consider correlation between residuals	-	Overall statistical test; supplementary to calibration graph	Very general
Subgroup calibration	Compare observed vs. predicted in subgroups	Table	Conceptually easy	Not sensitive to various miscalibration patterns

of predicted probabilities is the number of expected outcomes; this expected number is compared with the observed number in the ten groups with a χ^2 test. In model development, this χ^2 test has eight degrees of freedom; at external validation the degrees of freedom is 9. There are many drawbacks to the H-L test.[198,174] First, there are some technical issues: Should we always use deciles of predictions, or make the quantiles dependent on the sample size? Can we group by risk-interval, e.g. 0–10%, 11–20%, etc ("interval grouping")? Second, the test has poor power to detect miscalibration in the common form of systematic differences between outcomes in the new data and the model development data, or to detect overfitting of the effects of predictors. Some proposed that the H-L test should only be used in model development, in addition to more specific tests on model assumptions, such as tests for linearity (adding non-linear transformations) and additivity (adding interaction terms). Reported H-L tests are usually non-significant if they reflect apparent validation on the data that were also used to construct the model. Such non-significant results may contribute to the face validity of a model as perceived by some readers, but have no scientific meaning.

Alternative goodness-of-fit tests have been proposed with better statistical properties, such as the Goeman-Le Cessie goodness-of-fit test.[250,141] It assesses the alternative hypothesis that any nonlinearities or interaction effects have been missed in a logistic regression model. Such neglected effects can be detected by looking for patterns in the residuals: Observations close to each other in covariate space, which deviate from the model in the same direction. The approach is to smooth the regression residuals and to test whether these smoothed residuals have more variance than expected under the null hypothesis, which occurs when residuals that are close together in the covariate space are correlated. The test statistic is a sum of squared smoothed residuals.

Another approach to goodness-of-fit is to study observed vs. expected outcomes in subgroups of patients. For example, we can assess the difference between observed vs. expected outcomes in males and females, or other subgroups of patients. If the effect of the subgroup is not well modelled, e.g. an interaction was missed, this might be reflected in this assessment. There are however more direct ways of assessing the influence of subgroup characteristics, as was discussed in Chap. 13 on model specification. So, this check for calibration is also more for face validity of the model and for convincing potential users than a serious check of calibration. Measures for assessment of calibration are compared in Table 15.6.

15.3.9 Calibration of Survival Predictions

For survival outcomes, formal tests similar to the H-L test are possible by comparison of observed K-M percentages with average predictions across groups of patients. Furthermore, we can study the distribution of Cox-Snell residuals, in a plot of the cumulative hazard vs. the residuals, which should form a straight line.[174]

**15.3.10 Example: Calibration in Testicular Cancer Prediction Model*

For the prediction model of residual mass histology, we plot the actual outcome vs. predicted for the development sample and the validation sample (Fig. 15.10). We include the distribution of predicted risks, such that discrimination can also be judged. The results by decile of predicted risk are shown in Table 15.7, to clarify the calculation of the Hosmer-Lemeshow statistic. Other tests for miscalibration included the overall test for calibration-in-the-large and calibration slope, and the Goeman–Le Cessie test, which were non-significant for model development and external validation (Table 15.8).

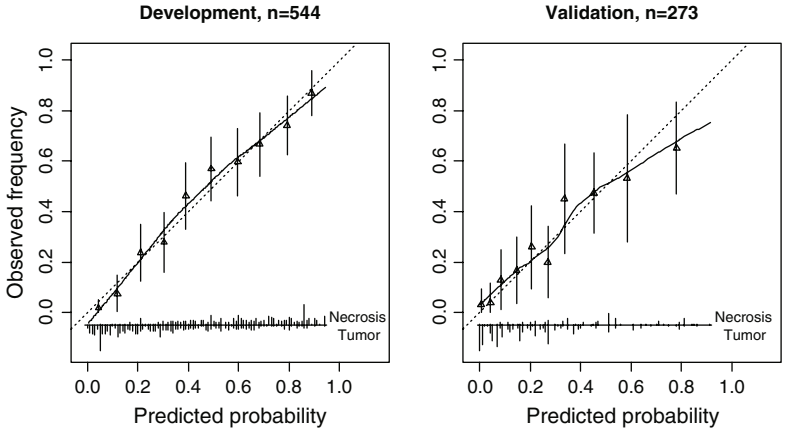

Fig. 15.10 Validity of predictions of necrosis in the development sample (n=544) and in the validation sample (n=273). The distribution of predicted probabilities is shown at the bottom of the graphs, separately for those with necrosis and those with residual tumor. The triangles indicate the observed frequencies by deciles of predicted probability

Table 15.7 Hosmer-Lemeshow test for calibration of the testicular cancer prediction model

		Development[a]				Validation[b]		
Decile	P(%)	N	Predicted	Observed	P(%)	N	Predicted	Observed
1	<7.3	56	2.4	1	<1.8	31	0.2	1
2	7.3–16.5	53	6.3	4	1.8–7.3	25	1.1	1
3	16.6–26.5	55	11.6	13	7.4–11.1	31	2.6	4
4	26.6–34.7	54	16.4	15	11.2–17.5	30	4.4	5
5	34.8–43.6	54	21.0	25	17.6–24.3	27	5.6	7
6	43.7–54.0	58	28.5	33	24.4–31.0	30	8.1	6
7	54.1–63.5	52	31.0	31	31.1–37.2	20	6.7	9
8	63.6–73.8	54	36.9	36	37.3–54.6	38	17.2	18
9	73.9–85.0	54	42.8	40	54.7–64.7	15	8.8	8
10	>85.0	54	48.0	47	>64.7	26	20.3	17
		544	245	245		273	74.9	76

[a]χ^2=5.9, df=8, p=0.66 [b]χ^2=9.2, df=9, p=0.42

Table 15.8 Calibration of testicular cancer prediction model

	Development	Internal validation	External validation
Calibration-in-the-large	0	0	–0.03
Calibration slope	1	0.97[a]	0.74
Calibration tests			
Overall miscalibration	p=1	–	p=0.13
Hosmer-Lemeshow	p=0.66	–	p=0.42
Goeman – Le Cessie[b]	p=0.63	–	p=0.94

Development and internal validation with n=544 patients, external validation in n=273 patients. Internal validation with 200 bootstrap resamples using Harrell's validate function

[a]Equivalent to the uniform shrinkage factor as discussed in Chap. 14

[b]Test statistics of squared smoothed residuals calculated with R program from Jelle Goeman, available from website

15.3.11 Calibration and Discrimination

The calibration plot can be extended into a "validation plot" as a central tool to visualize model performance. Calibration is shown by observed outcomes being close to prediction, while discrimination aspects can be indicated with the distribution of the predicted probabilities. The distribution can be shown by a histogram or density distribution. We can also make separate histograms for those with and without the outcome for further insights (see e.g. Fig. 15.10). It also helps to see the separation according to quantiles of predicted probabilities. For example, when deciles are used, these will be relatively far apart for a good discriminating model.

Calibration-in-the-large is a phenomenon that is fully independent of discrimination. For example, we can change the incidence of the outcome in a case-control study, but the discrimination will be unaffected. The calibration slope however has a direct relationship with discrimination. If the calibration slope is below unity, the discrimination is lower. Hence, overfitted models will show both poor calibration and poor discrimination when validated in new patients (Chap. 19).

Perfect calibration is possible with poor discrimination, for example when the range of predicted probabilities is small, such as between 9 and 11% for an average incidence of the outcome of 10%. At external validation, such a small range in predictions may arise from a narrow selection of patients (homogeneous case-mix). A drop in discriminative ability compared with the development setting can hence be explained by overfitting (calibration also poor), or a more homogeneous in case-mix (independent of calibration, see Chap. 19). On the other hand, a well discriminating model can have poor calibration, which can be corrected with various updating methods (Chap. 20).

*15.3.12 R Code

The Hosmer-Lemeshow test is implemented in a simple function `hl.ext` at the book's website. The user can specify the number of groups (ten by default) and degrees of freedom (groups – 2 for model development, groups – 1 for model validation).

Calibration plots are made by an extension of Harrell's `val.prob` function, called `val.prob.ci`. This function also provides assessment of calibration-in-the-large, calibration slope, and the calibration test *p*-value. Goeman provided R code for the functions mlogit (for binary of multinomial logistic regression), `smoothU` (for calculation of smoothed residuals), and `testfit` (for the Goeman-Le Cessie goodness-of-fit test).

15.4 Concluding Remarks

In this chapter we have discussed a number of performance measures for prediction models; many more can be used, as systematically discussed in work by Hilden, Bjerregaard, and Habbema in the 1970s.[161,162,163,191,192] Many performance measures are related to each other; e.g. the *c* statistic is related to the Mann-Whitney U statistic,

which is calculated as a rank order test for the difference between predictions by outcome. The *c* statistic is also linearly related to Somer's *D* statistic ($c = D/2 + 0.5$).

From a simple statistical perspective we want a small distance between observed outcome Y and predicted outcome $\hat{Y}$. Explained variation (R^2) can then be used to indicate performance, and indicates the predictability of the outcome: How much do we know already about the phenomena that lead to the outcome.[372] Diagnostic prediction models would hence be expected to have higher R^2 than prognostic models with long-term outcome. Indeed, prognostic models usually have R^2 around 0.20. This indicates that substantial uncertainty remains at the individual level; we can only provide probabilities, and no certainty on the individual outcome.[13,112]

We have focused on measures that are in wide use in medical journals nowadays, including the concordance statistic ('*c*', or area under the ROC curve) for discrimination, and various tests for calibration and goodness-of-fit. The *c* statistic has been criticized by some, and should not be the only criterion in assessment of model performance. Especially, *c* may be rather insensitive to inclusion of additional predictors in prediction models, such as novel biomarkers.[79,330] But our theoretical examples and case study show that the *c* statistic is a key measure; it is closely related to other performance measures such as R^2 and Brier score.

In principle we might focus our modelling strategy on optimizing performance measures such as the *c* statistic. Indeed, estimation algorithms have been described that maximize the *c* statistic rather than the log likelihood.[332]

Compared with current practice, calibration should receive more attention when evaluating prediction models. The recalibration test and its components (calibration-in-the-large and calibration slope) should be used routinely in performance assessment in external validation of prediction models.

15.4.1 Bibliographic Notes

The framework of a recalibration model was already proposed by Cox,[86] and has been supported by many other researchers for evaluation of model performance.[81,174,290,291,458] Nice illustrations of diagnostic test evaluation with ROC curves are available at:
http://www.anaesthetist.com/mnm/stats/roc/
Nice illustrations of Lorenz curves and the Gini index are at:
http://en.wikipedia.org/wiki/Gini_coefficient

Questions

15.1 Overall performance measures
Overall performance measures for logistic regression models include Brier score and R^2 type of measures, such as Nagelkerke's R^2.
(a) What values can Brier scores and R^2 take?
(b) What types of scoring rules are Brier and R^2?
(c) What are disadvantages of Brier and R^2?

15.2 Lorenz curve and incidence (Fig. 15.6)
In a Lorenz curve, the visual impression of a model with a c statistic of 0.80 depends on the incidence of the outcome.
(a) What happens when a Lorenz curve is made for situation with 1% incidence?
(b) And what for 99% incidence?

15.3 Interpretation of validation graph (Fig. 15.10)
Validity of predictions can well be judged graphically. How do you judge
(a) calibration-in-the-large?
(b) calibration slope?
(c) discrimination?

15.4 Relationship between calibration, discrimination, and overall performance.
Explain the differences and the relation between calibration, discrimination, and overall performance measures.