Table 15.7 A hierarchy of calibration levels for risk prediction models for binary outcomes [602]

Strength of calibration	Definition	Assessment
Mean	Observed event rate equals average predicted risk	Compare event rate with average predicted risk; Evaluate calibration-in-the-large as $a\|b = 1 = 0$, 1 *df* test
Weak	No systematic overfitting or underfitting and/or overestimation or underestimation of risks	Calibration analysis for calibration-in-the-large and calibration slope; evaluate with Cox recalibration test: 2 *df* test of the null hypothesis that $a = 0$ and $b = 1$ [114]
Moderate	Predicted risks correspond to observed event rates	Calibration plot with smooth curve, and/or inspection by grouped predictions
Strong	Predicted risks correspond to observed event rates for each and every covariate pattern	Scatter plot of predicted risk and observed event rate per covariate pattern, impossible with continuous predictors

15.3 Calibration

Another key property of a prediction model is calibration, i.e., the agreement between observed outcomes and predictions. The most common definition of calibration is that if we observe *p*% risk among patients with a predicted risk of p%. So, if we predict 70% probability of residual tumor tissue for a testicular cancer patient, the observed frequency of tumor should be approximately 70 out of 100 patients with such a predicted probability. Weaker forms of calibration only require the average predicted risk (mean calibration) or the average prediction effects (weak calibration) to be correct. Strong calibration requires that the event rate equals the predicted risk for every covariate pattern [602]. This implies that the model is fully correct for the validation setting (Table 15.7). Graphical inspection is very useful (a calibration plot) [26, 225].

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 at 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. Such probabilities are not observed directly. 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.,

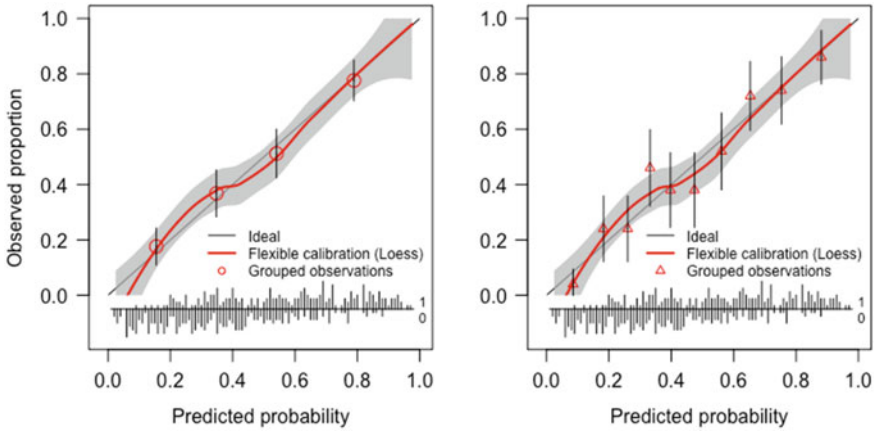

Fig. 15.11 Calibration plots of actual outcome versus predictions for a hypothetical model with *c* statistic 0.76, $n = 500$. Left and right panels only differ in the number of groups (4 vs 10). The distributions of actual 0 and 1 values are shown at the bottom of the graph; the *loess* smoother (with 95% confidence band) is close to the ideal 45 degree line; actual outcomes according to risk groups are shown by circles and triangles (each circle: $n = 125$; triangle: $n = 50$)

using the *loess* algorithm [26]. Confidence intervals can be calculated for such a smooth curve (Fig. 15.11).

The plot visualizes mean calibration (are observed outcome systematically lower or higher than predicted?), weak calibration (is there a general trend in predictions being too extreme, a sign of overfitting at model development?). Moreover, we can plot results for subjects grouped by similar probabilities. This allows us to assess a moderate level of calibration by comparing mean observed proportions per group to the mean predicted outcome. For example, we can plot observed outcome in groups defined by quintile or by decile of predictions (Fig. 15.11). This makes the plot a graphical illustration of the Hosmer–Lemeshow goodness-of-fit test. Note that we also learn about discrimination from a calibration plot: A better discriminating model has more spread between observed proportions per group than a poorly discriminating model. The choice of groups is important for the visual impression of calibration; if small groups are plotted, the variability will be larger (right panel in Fig. 15.11).

15.3.2 Mean and Weak Calibration at Internal and External Validation

The mean calibration will usually be perfect when we compare observed outcomes to the mean predictions in the data set used to develop a model. Such apparent calibration is hence not informative. Similarly, the mean calibration remains

uninformative with internal validation techniques such as cross-validation or bootstrapping (see Chap. 17). In contrast, when we validate the model in external data, the calibration of the mean risk is often far from perfect ("poor calibration-in-the-large").

The concept of weak calibration 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., log odds (or logit) in logistic regression. The $b_{overall}$ is named the calibration slope [114]. Ideally, the calibration slope $b_{overall} = 1$. With apparent validation, $b_{overall} = 1$, since 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 ($b_{overall} < 1$) [109, 627]. 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 phenomena: overfitting on the development data and true differences in the effects of predictors.

15.3.3 Assessing 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 a model with the residual $y_{new} - \hat{y}$ as the outcome. 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:

$$\begin{aligned} \text{OR} &= \text{odds}(\text{mean}(\hat{y}))/\text{odds}(\text{mean}(y_{new})) \\ &= [\text{mean}(\hat{y})/(1 - \text{mean}(\hat{y})]/[\text{mean}(y_{new})/(1 - \text{mean}(y_{new})]. \end{aligned}$$

For statistical testing of the difference, we need to be more careful. In logistic regression, the relation between the outcome y and the linear predictor is nonlinear (i.e., logistic). We want to compare logit(y_{new} = 1) to logit($\hat{y}$), where

$$\text{mean}(\text{logit}(y_{new} = 1) - \text{logit}(\hat{y})) \text{ is not equal to } \text{mean}(\text{logit}(y_{new} = 1)) - \text{mean}(\text{logit}(\hat{y})).$$

In a model, we could write

$$\text{logit}(y_{new} = 1) - \text{logit}(\hat{y}) = a; \text{ or}$$
$$\text{logit}(y_{new} = 1) = a + \text{logit}(\hat{y}) = a + \text{offset(linear predictor)}.$$

The intercept a then reflects the difference in log odds between predictions and observed outcome, adjusted for the linear predictor. The offset makes that predictions are taken literally, as in linear regression. 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 alternative hypothesis is a <> 0 | $b_{overall}$ = 1 (Table 15.8).

Note that exp(a) can be interpreted as an observed-to-expected (O/E) ratio. This ratio can also be calculated directly by comparing the sum of observed events (O) with the sum of the predictions (E). These ratios will differ, with exp(a) larger than the simple O/E ratio [210]. This is because the estimation of a was conditional on the linear predictor (as an offset variable), which makes for an adjusted estimate rather than an unadjusted estimate as for O/E. So, exp(a) can be interpreted as the odds ratios for individuals, given their covariate pattern (a conditional estimate), while O/E reflects the overall average miscalibration (a marginal estimate).

The calibration slope can be estimated from the recalibration model logit(y_{new} = 1) = a + $b_{overall}$ * logit($\hat{y}$) = a + $b_{overall}$ * linear predictor. The deviation of the calibration slope from 1 ("miscalibration") can be tested by a model that includes an offset variable:

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

The slope coefficient $b_{overall}$ reflects the deviations from the ideal slope of 1, and can be tested with Wald or LR statistics (Table 15.8).

For a survival outcome, the calibration slope $b_{overall}$ can be assessed as

Table 15.8 Calibration tests for prediction model $y \sim a + b_{overall} * \hat{y}$. H_0 and H_1 indicate the null and alternative hypothesis, respectively

	H_0	H_1	*df*
Calibration-in-the-large	a = 0 | $b_{overall}$ = 1	a <> 0 | $b_{overall}$ = 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

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

The model for deviation from a slope of 1 is

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

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

This recalibration test for $a = 0$ and $b = 1$ has several advantages. 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).

With a parametric survival model, we can specify parameters which reflect differences in average survival, after adjustment for predictor effects. Van Houwelingen transformed the baseline hazard from a Cox model to a Weibull model [623]. 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 a single coefficient for the linear predictor, to estimate a recalibrated model.

15.3.4 Assessment of Moderate Calibration

Moderate calibration can best be assessed graphically as discussed above with the calibration plot (Sect. 15.3.1). Smooth curves can be constructed with the *loess* smoother, which may be considered nonparametric, or with a spline smoother such as a restricted cubic spline [26, 602]. A formal test might be done for nonlinearity: compare the fit with an *rcs* versus a linear offset term. In addition to graphical inspection of grouped predictions, we may perform a chi-square test for observed versus expected numbers by group. This test is the often used Hosmer–Lemeshow (HL) test [257].

For the HL test, patients are grouped typically by decile of predicted probability. The sum of predicted probabilities is the number of expected outcomes; this expected number is compared to the observed number in the 10 groups with a chi-square test. At model development, this chi-square test has eight degrees of freedom; at external validation the degrees of freedom are nine. There are many drawbacks to the H-L test [225, 256]. First, there flexibility in the grouping: should we always use deciles to group predictions in tenths, or make the quantiles dependent on the sample size? Should 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 nonlinear transformations) and additivity (adding interaction terms). Reported H-L tests are usually nonsignificant if they reflect apparent validation on the data that were also used to construct the model. Such nonsignificant results may contribute to the face validity of a model as perceived by some readers, but have no scientific meaning.

Various other measures are available for moderate calibration. An intuitively appealing measure of calibration is the absolute difference between smoothed observed outcomes and predicted probabilities (Harrell's E statistic) [225]. 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 well be judged visually in a calibration plot such as Fig. 15.11, with the distribution of predictions at the x-axis. We can also summarize the miscalibration in a single index such as the estimated calibration index (ECI) [621], which also summarizes the distance between a smooth calibration curve and the ideal 45° line.

15.3.5 Assessment of Strong Calibration

The concept of strong calibration is related to goodness-of-fit, which relates to the ability of a model to fit a given set of data. Ideally, we would identify a true underlying model, which may be utopic (Chap. 5) [602]. Typically, there is no single goodness-of-fit test which has good power against all kinds of lack of fit of a prediction model. Examples of lack of fit are missed nonlinearities, interactions, or an inappropriate link function between the linear predictor and the outcome. Such deviations are better assessed by adding nonlinear terms to a model, adding interaction terms, and examining alternative link functions, if sample size allows for such flexibility in modeling strategy.

An interesting approach is the Goeman–Le Cessie goodness-of-fit test [187, 324]. 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 studying 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. This deviation 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 versus expected outcomes in subgroups of patients, defined by predictor values. For example, we can assess the difference between observed versus expected outcomes in males and females, or other subgroups of patients. If the effect of the subgroup is not well modeled, 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 summarized in Table 15.9.

15.3.6 Calibration of Survival Predictions

In a survival context, we can assess calibration-in-the-large with a model-based approach [116]. This involves using a Poisson model which uses the linear predictor as an offset. Additionally, we can fit the linear predictor based on the original prediction model as a predictor to obtain the calibration slope. Furthermore, we can fit a model with the linear predictor as an offset and adjusting for a set of dummy variables created by deciles of the linear predictor from the original prediction model. A score test for the group effect of this set is asymptotically equivalent to the Grønnesby and Borgan, or Nam–D'Agostino tests, which are survival analysis variants of the Hosmer–Lemeshow test [132]. Again, these tests produce a p-value that is difficult to interpret: with small external validation samples, we lack statistical power to detect miscalibration. On the other hand, we will commonly find statistically significant miscalibration with large external validation samples. These tests are therefore not very useful.

A calibration plot can also be produced. The calibration of a model can be studied at fixed time points. We can group patients for calculation of survival rates with the Kaplan–Meier method. Harrell suggests to use at least 50 subjects per group, depending on the hazard of the outcome [225]. This observed survival may be compared to the mean predicted survival from the prediction model. A smoothed calibration curve can be obtained by comparing Cox–Snell residuals on the cumulative probability scale against the right-censored survival times [225]. We can also plot the observed t-year risk of the outcome for each tenth of patients (and 95% confidence intervals) against the predicted risk estimated from the Poisson regression model [116]. This model-based approach can be extended to replace the groups with splines. These approaches depend on the baseline hazard being available either for at least some specific time points [471].

15.3.7 Example: Calibration in Testicular Cancer Prediction Model

For the prediction model of residual mass histology, we plot the actual outcome versus predicted for the validation sample (Fig. 15.12). We include the distribution of predicted risks, such that discrimination can also be judged. The results for five

Table 15.9 Summary of measures for calibration of a prediction model for binary outcomes

Performance aspect	Calculation	Visualization	Pros	Cons
Calibration-in-the-large	Compare mean (y) versus 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 versus $\hat{y}$	Calibration graph	Conceptually easy, summarizes miscalibration over whole curve	Depends on smoothing algorithm
Hosmer–Lemeshow test	Compare observed versus 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 versus predicted in subgroups	Table	Conceptually easy	Not sensitive to various miscalibration patterns

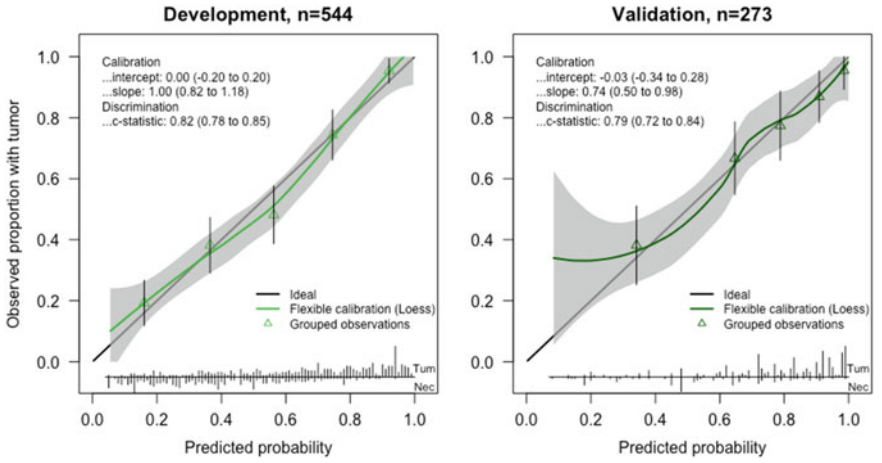

Fig. 15.12 Validity of predictions of tumor in the testicular cancer development sample (n = 544) and in the validation sample (n = 273). The distribution of predicted probabilities is shown at the bottom of the graphs, separately for those with tumor and those with necrosis ("Tum" vs. "Nec"). The triangles indicate the observed frequencies by tenth of predicted probability

Table 15.10 Calibration of testicular cancer prediction model[a]

	Development	External validation
Calibration-in-the-large	0	–0.03
Calibration slope	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]	$p = 0.63$	$p = 0.94$

[a]Internal validation with 500 bootstrap resamples using Harrell's `validate` function
[b]Test statistics of squared smoothed residuals calculated with an R program from Jelle Goeman, available at www.clinicalpredictionmodels.org

risk groups of predicted risk are shown. Tests for miscalibration included the overall test for calibration-in-the-large and calibration slope, and the Goeman–Le Cessie test, which were nonsignificant for model development and external validation (Table 15.10). Note that assessment of calibration makes little sense in the development data, while it is essential at external validation.

*15.3.8 *R Code for Assessing Calibration*

Calibration plots were made by an extension of Harrell's `val.prob` function called `val.prob.ci.2` [602]. This function also provides assessment of

calibration-in-the-large, calibration slope, and the calibration test *p*-value. Goeman developed 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).

15.3.9 Calibration and Discrimination

The calibration plot can be extended into a "validation plot" as a central tool to visualize model performance [568]. Moderate 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., Figs. 15.10 and 15.12). It also helps to see the separation according to quantiles of predicted probabilities. For example, when deciles are used to define tenths, 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 mathematical relation with discrimination [629]. If the calibration slope is below unity, the discrimination is also lower at external validation. Hence, overfitted models will show both poor calibration and poor discrimination when validated in new patients (Chap. 19).

Weak to strong 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 to the development setting can hence be explained by overfitting (calibration poor), or a more homogeneous case-mix (independent of calibration, see Chap. 19) [629]. On the other hand, a well discriminating model can have poor calibration, which can be corrected with various updating methods (Chap. 20).

15.4 Concluding Remarks

In this chapter, we have discussed a number of performance measures for prediction models; many more can be used, as already systematically discussed in work by Hilden, Bjerregaard, and Habbema in the 1970s [215–217, 247, 248]. 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 Somers'

D statistic ($c = D/2 + 0.5$). Recently proposed measures for reclassification have many links to more traditional measures [428, 429].

From a simple statistical perspective, we want a small distance between observed outcome y and predicted outcome $\hat{y}$. Explained variation (R^2) can be used to indicate performance, and quantifies the predictability of the outcome: how much do we know already about the phenomena that lead to the outcome [491]? Diagnostic prediction models would hence be expected to have higher R^2 than prognostic models with long-term outcome. Indeed, prognostic models usually only have R^2 around 20–30%. This indicates that substantial uncertainty remains at the individual level; we can only provide probabilities, and we are far away from providing certainty on the individual outcome [13, 150].

We have focused on measures that are in wide use in medical research nowadays, including the concordance statistic (c, or area under the ROC curve, AUC) for discrimination, and various tests for calibration and goodness-of-fit. We gave some attention to Lorenz curves, although these are not often used; we did not discuss predictiveness curves, which provide useful insight in some applications [432]. The c statistic has been criticized by some, and should not be the only criterion in assessment of model performance. Especially, c is considered to be rather insensitive to inclusion of additional predictors in prediction models, such as novel biomarkers [107, 426]. 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 [434]. Improvements in model fit will also show improvements in c statistic.

In principle, we might focus our modeling 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 [431].

Compared to current practice, calibration should receive more attention, especially when externally validating prediction models [103]. The recalibration test and its components (calibration-in-the-large and calibration slope, with recalibration parameters a and b) should be used routinely in performance assessment at external validation of prediction models.

15.4.1 Bibliographic Notes

The framework of a recalibration model was already proposed by Cox [114], and has been supported by many other researchers for evaluation of model performance [109, 225, 379, 380, 626]. Nice illustrations of diagnostic test evaluation with ROC curves are available at: http://www.anaesthetist.com/mnm/stats/roc/ and illustrations of Lorenz curves and the Gini index are at: http://en.wikipedia.org/wiki/Gini_coefficient.