

Chapter 9

General discussion

Clinical prediction models can be helpful tools in decision making. They provide the probability that the disease of interest is present (diagnosis) or that a particular health state will occur during a disease process (prognosis). A high predicted probability of the diagnostic or prognostic outcome supports the diagnosis or the decision to offer a particular treatment in order to improve the prognosis. Before a prediction model can be applied, the performance of the model in new patients (external validity) needs to be studied.^{1,2} This thesis described several aspects of model validity and the behaviour of a number of relevant performance measures. Further, the measures were applied to study the external validity of a prediction model for residual mass histology in testicular germ cell cancer. The focus was mainly on models developed with logistic regression analysis. In this chapter, the specific objectives as defined in the introduction will be discussed. The objectives were:

1. To describe aspects of validity and relevant performance measures for clinical prediction models;
2. To estimate the power of performance measures to detect poor validity;
3. To externally validate a prediction model for residual mass histology in testicular cancer patients;
4. To update a prediction model for residual mass histology in testicular cancer patient.

Aspects of model validity and relevant performance measures

In chapter 2, three aspects of model validity were discussed: calibration; discrimination; and clinical usefulness. For each aspect several performance measures were described including graphical measures, summary measures, and statistical tests.

Ideally, all three validity aspects are found to be perfect. However, one aspect may be more relevant than another given the clinical purpose for which the model was developed. Two important purposes in clinical practice are:¹

1. to inform the patient or the patients' family about a diagnosis or a prognosis;
2. to support the diagnostic process or to decide which treatment the patient should receive to improve the prognosis.

If the model is used to inform the patient, it is particularly important that the model provides well-calibrated predictions of the outcome. Even a correct estimate of the average disease prevalence may be valuable. For instance, knowing that the average risk of having metastasis is 30% in clinical stage I testicular germ cell cancer may be more reassuring than only knowing to have a substantial risk.

Using the prediction model as a decision tool, like in this thesis, the ability of the model to discriminate between a patient with the outcome (e.g. a benign residual mass after chemotherapy) and a patient without the outcome (e.g. residual tumour) is important. Discrimination refers to a correct ranking of predicted probabilities, i.e. patients without the outcome have low predicted probabilities, patients with the outcome have high predicted probabilities. In clinical practice however, a threshold value is required to make treatment choices. Measures of clinical usefulness take this threshold value into account. Clinical usefulness implies that application of the model at a particular threshold value can improve the treatment choices as compared with a reference policy. The threshold value is defined by the relative seriousness of the consequences of false-positive (classified diseased, while not-diseased) and false-negative classifications (classified non-diseased, while diseased). These consequences should also be considered. Therefore, clinical usefulness measures are closely related to utility-based measures that consider the patients benefit or loss as a consequence of a certain decision.³

A more accurate measure of clinical usefulness: change in weighed error rate

In chapter 2, a measure for clinical usefulness was described that classified patients and took into account the relative seriousness of false-positive and false-negative classifications. The total number of falsely classified patients was estimated at a threshold value of 70%. This threshold value implies that false classifications of benign tissue leading to unresected tumour should be weighed 7/3 times higher than false classifications of residual tumour leading to unnecessarily resected masses. The relative difference in the weighed number of false classifications, expressed as ‘unnecessarily resected masses’, between the reference policy and the prediction model was used as a measure of clinical usefulness.

This measure has a very atypical unit of ‘unnecessarily resected masses’. A more general measure would be a weighed error rate, which is unitless. The unweighed error rate is the total number of misclassifications divided by the total number of subjects, and equals 1-accuracy. The possible values range from 0 to 1. The weighed error rate ranges also from 0 to 1 and takes into account the differences in severity of false-positive and false-negative classifications. For the residual mass histology classification, the weighed error rate at a threshold value of 70% may be defined as:

$$(0.3*FP + 0.7*FN) / (0.3*benign + 0.7*tumour)$$

with FP: falsely classified as tumour; FN: falsely classified as benign

The difference in the weighed error rates of the reference policy and the prediction model is a measure for clinical usefulness. The decrease in weighed error rate was 0.01 (on a scale from -1 to 1) for the data from Indiana University Medical Center (IUMC), which was very small and similar to the decrease in weighed misclassifications estimated in chapter 2, which was 1%.

The change in weighed error rate could also be estimated for the other data (development data and first validation [chapter 5] and the EORTC/MRC trial [chapter 7]). For the patients of the development data, application of the prediction model would have decreased the weighed error rate from 0.28 for the reference policy to 0.20 for the model. 26% of the predicted probabilities were above the threshold value. Even 30% of the patients from the first validation had predicted probabilities above the threshold value. Nevertheless, the weighed error rate decreased only from 0.28 to 0.24. The predicted probabilities between 70% and 90%, which were in general too high (Figure 5.2B), negatively influenced the clinical usefulness for these patients. Nearly all patients derived from the EORTC/MRC trial (96%) had predicted probabilities below the threshold value. The decrease in weighed error rate was therefore small. Since the few patients with predicted probabilities above the threshold value were correctly classified as having benign tissue the decrease in weighed error rate was 0.01 instead of exactly 0.

Theoretical combinations of the three validity aspects

We showed that a model with reasonable calibration and good discrimination not automatically resulted in a clinically useful model (chapter 2). Here, several other combinations of calibration, discrimination, and clinical usefulness are shown. To study the different combinations, samples were generated analogously to the methods described in chapter 3.

A validation plot of a model with all three aspects qualified as ‘good’ is shown in Figure 9.1A. The line shows perfect agreement between the observed frequencies and the predicted probabilities, with intercept = 0 and slope = 1. The discriminative ability is good, i.e. c -statistic is 0.81. The weighed error rate changed from 0.33 for the reference policy to 0.22 for the model at a threshold value of 70%. An example of a model with well-calibrated predictions and good discriminative ability, but poor clinical usefulness is shown in Figure 9.1B. The predictions are correct and discriminate well (c -statistic=0.81), but all predictions are below the threshold value leading to the same decisions as with the reference policy. This figure also shows that the clinical usefulness of the model can never be lower than the reference policy as long as the predictions are correct. A more counterintuitive example is shown in Figure 9.1C. The model has low discriminative ability (c -statistic = 0.58), but since the predictions are correct and many are above the threshold value, the model is clinically useful (decrease in weighed error rate of 0.13). Poor calibration as shown in Figure 9.1D (predictions on average too high with intercept=-0.71) does not necessarily result in a useless model. A decrease in weighed error rate was found, although small (0.03). However

carefulness is required. Extreme poor calibration (for instance intercept=-1.0) may worsen the decision making process (change in weighed error rate=- 0.02).

In conclusion, Figure 9.1 shows that adequate calibration and good discrimination are not sufficient to improve clinical decision making. The predicted probabilities may show a reasonable spread, leading to a *c*-statistic over 0.80, but only few decisions will be influenced if the largest part of the predictions are below the threshold value. Counterintuitive combinations, as shown in Figure 9.1C and 9.1D, may be interesting for further research.

The listing of the performance measures in chapter 2 was not extensive. More research may be performed on overall measures such as the Brier score and R^2 and the relation between calibration and discrimination measures with measures for clinical usefulness. Further, it was shown that the discriminative ability of a prediction model depends on the case-mix in the validation data. Hence, the *c*-statistic is not only a model characteristic. The influence of the case-mix on the discrimination may be further studied with simulation data.

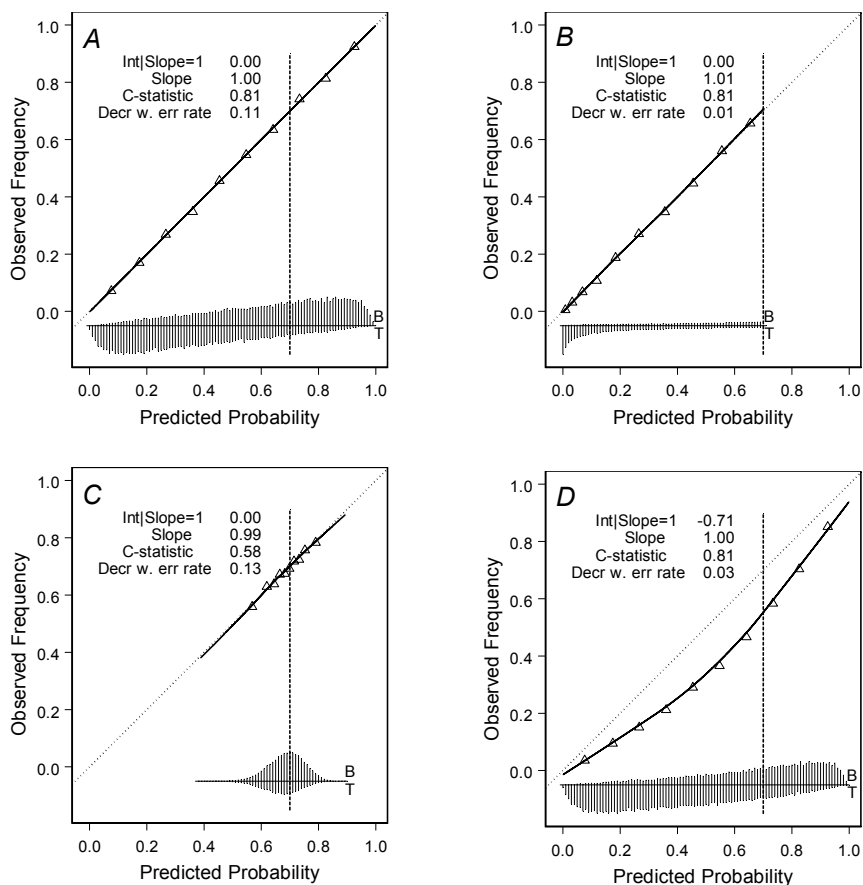


Figure 9.1. Four combinations of calibration; discrimination; and clinical usefulness shown in validation plots. A: all three aspects are good; B: despite good calibration and discrimination, low clinical usefulness; C: calibration good, despite poor discrimination, high clinical usefulness; D: predictions systematically too high, discrimination good, and still some clinical usefulness.

The power of performance measures to detect lack of validity

In chapter 3, we studied the power of several performance measures to detect differences in model performance. Further, required effective sample sizes (i.e. number of events) were estimated. Samples with at least 100 events and 100 non-events were recommended to achieve adequate power to detect possible poor validity.

Unfortunately, the samples used in the chapters 5-7 to validate the prediction model for residual mass histology contained all less than 100 events. The numbers of events were 77, 76, and 27 in the first validation study, the IUMC data, and the EORTC/MRC trial, respectively. The intercept of -0.19 for the IUMC data might perhaps have become statistically significant different from 0, if the sample had contained more events. That would have implied that the predicted probabilities for the IUMC patients were systematically too high and adjustment of the model intercept might have been reasonable. However, the current study did not contain enough evidence to support such a recommendation.

The poor calibration found in the patients from the EORTC/MRC trial may be simply a result of the low number of patients ($n=105$). Only a few poor predictions were sufficient to distort the visual impression of calibration. For illustration, four random samples of 105 patients (27 events and 78 non-events) were drawn from the development dataset. Validating the model, that was by definition valid in the complete dataset, also showed poor graphical calibration in the small samples (Figure 9.2).

In conclusion, sample size recommendations were formulated based on a simulation study with one particular prediction model and a limited number of differences in model performance. Further empirical research with other prediction models and more subtle differences in model performance may refine the current recommendations. The mathematical relations between the effective sample size and the power of measures to detect a difference in performance may be used in further research to propose a table listing the required number of events to detect relevant differences in model performance.

External validity of a prediction model for residual mass histology

The external validation studies of a prediction model for residual mass histology in advanced testicular germ cell cancer described in the chapters 5-7 showed all reasonable calibration (see Figures 5.2, 6.1B, and 7.2). The calibration slope was for two of the three studies close to 1 (0.97, 0.91, and 0.70 for the first validation study, the IUMC data, and the EORTC/MRC trial) and in none statistically significant different from 1. The slope estimates were all smaller than 1, which may be an indication for insufficient shrinkage of the regression coefficients in the development process. The intercept of the calibration line was in two studies smaller than 0 (-0.20, -0.19, and 0.02 respectively), though again not statistically significant. An intercept different from 0 would imply that the patients in the validation data were different from the patients used for the development in a way which was not reflected in the model parameters.

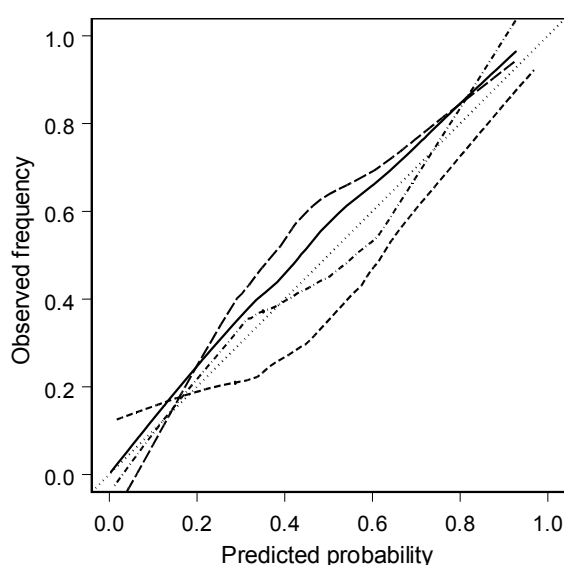


Figure 9.2 Miscalibration in four random samples. Per sample 105 patients were drawn from the development data, illustrating the variability in validity that may occur by change only.

We may think of a missed predictor in the model, for example related to the referral pattern of patients.

The discriminative ability of the model was good or reasonable, with c -statistics of 0.82, 0.79, and 0.76 (Table 9.1). However, discrimination is not only a model aspect, but also a population aspect. A statistically correct model will have less discriminative ability in a more homogeneous population. For a better interpretation of the c -statistic, the spread of predicted probabilities in the validation data was studied (case-mix) given that the predictions were correct. For each validation dataset a sample was simulated of 5000 patients having the same covariate patterns as the patients in the validation data. The outcome values were generated under the assumption that the regression coefficients of the prediction model were correct. The c -statistics estimated in the simulated samples were 0.84, 0.80, and 0.82, respectively. These values indicate how much discrimination could be realised in the validation data given that the model was correct. The discrepancy between the estimated discrimination and the theoretically realisable discrimination was highest for the patients from the EORTC/MRC trial, which may be coincidence due to the small sample size in this study.

Like other measures, measures for clinical usefulness are also optimistic when estimated in the development data. In the validation data the decrease in weighed error rate was always smaller than the 0.08 as estimated for the development data: 0.04, 0.01, and 0.01, respectively. However, the low clinical usefulness in the validation data may also be a result of a change in case-mix. Similar to the simulations to estimate the theoretically realisable c -statistic, the theoretically realisable decrease in weighed error rate was estimated. A model with correct regression coefficients will not be more useful for the patients from the EORTC/MRC trial. In contrast, a substantial gain in clinical usefulness may be achieved for the patients from the first validation study.

Table 9.1 Empirical estimates versus theoretically realisable estimates given correct predicted probabilities of discrimination (c-statistic) and clinical usefulness (decrease in weighed error rate) in the validation data

<i>Data</i>	<i>C-statistic</i>		<i>Decrease in weighed error rate</i>	
	<i>Empirical</i>	<i>Theoretical upperbound</i>	<i>Empirical</i>	<i>Theoretical upperbound</i>
First validation	0.82	0.84	0.04	0.11
IUMC	0.79	0.80	0.01	0.03
EORTC/MRC trial	0.76	0.82	0.01	0.01

Practical aspects of validation studies

To validate a prediction model in new patients, information on a number of variables is required.⁴ The values of the predictors in the model are used to calculate the predicted probabilities, which are then compared with the observed outcome values. If all this information is available, the model validity can be assessed.

The analyst can already anticipate on the applicability of the model during the development process. Variables that are routinely measured can be selected instead of rarely determined parameters. Although the model for residual mass histology contains readily available predictors, many values for the serum levels of lactate dehydrogenase (LDH) were missing for the IUMC patients (chapter 6). LDH has relatively recent been identified as a predictor and may therefore be missing for patients treated longer ago. If the proportion of missing values is large, imputing the values may not be meaningful. The data contain then to few information on the correlation with other covariates to derive sensible imputation values. Deleting all cases with missing values is never a good option. This may result in biased estimates and lower precision.⁵ In our case, exclusion of patients with missing values would leave only 74 of the 276 patients. Therefore, we developed an alternative model without LDH using the development data and performed a validation study on the IUMC patients with the alternative model.

Another data problem may be differences in documentation. For instance, variables can be documented categorically instead of continuously, or a variable definition can differ between centres or in time. In the IUMC data, mass size after chemotherapy and change in mass size during chemotherapy were categorically documented, while the variables were continuously entered in the model. A mean value for each category might have been estimated, derived for instance from the development data (if available, as in our case). The means for the different categories could have been filled in and application of the model would have been possible. Since we already had to develop an alternative model for the missing values of LDH, we chose to include categorical variables in the alternative model for postchemotherapy mass size and change in mass size during chemotherapy. This model could be validated in the IUMC data.

A definition problem was found with the systematic review for predictors of occult metastasis (chapter 4). Different classifications were used for teratoma histology. The British classification scheme divides teratoma in teratoma differentiated, mature teratoma intermediate, and mature teratoma undifferentiated, while most other European countries and the US divide teratoma in ‘mature’ and ‘immature’. The absence of teratoma differentiated and the absence of mature teratoma are both identified as predictors for occult metastasis. The predictive strength could be quantified for each definition. However, to develop a model that includes all relevant predictors, one definition should be chosen. The more widely used definition will be most appropriate, although this may give problems for application in British patients.

External validity of prediction models in a broader perspective

The aim of an external validation study is to find evidence for the generalisability of the prediction model to another setting. However, the results of such a study are very specific for the model and setting under study. Results of particular validation studies as shown in the chapters 5-7 are therefore difficult to generalise. The model for residual mass histology was valid for patients from a German centre that was included in the first validation study, but how about other German centres or a centre in France? The prediction model may need an intercept adjustment for the patients from IUMC and possibly also for the patients from Memorial Sloan Kettering Cancer Center. Does that mean that the model should be adjusted for all US centres, or only for these highly specialised US centres? Unfortunately, the described studies can not fully answer these questions.

The validation studies had in common that the odds ratios for the predictors in the model were similar to the odds ratios estimated in the development data. This is in agreement with a more general idea that biological associations do not change much over time or between centres. For instance, elevated serum levels of alpha-fetoprotein (AFP) before chemotherapy will for all patients, independent of place or time, indicate that the metastatic lesion contains yolk sac elements. Chemotherapy will affect the metastatic lesions containing yolk sac similarly in different centres. Therefore, the regression coefficient of elevated AFP will also be similar for those patients.

In contrast to such biological associations, the intercept of the model is more sensitive to variations in time and place. A different intercept can be found if the average outcome frequency differs in the new data while the case-mix, as defined by the predictors in the model, is similar. This implies that the change in the outcome frequency cannot be explained by the predictors in the model. In a population with a different average outcome frequency, but also a different case-mix, the case-mix may completely explain the difference in frequency. In that case, the calibration can still be good. In contrast, the discriminative ability of the model will be affected, because a difference in case-mix may reflect a difference in population heterogeneity, which influences the discriminative ability of the model.

In the example of the association between the serum level of AFP and presence of yolk sac elements, a subgroup effect may influence the individual regression coefficient of AFP. For instance, a new chemotherapeutic agent that would be particularly effective for yolk sac elements, will change the probability that metastases with yolk sac elements are successfully treated and therefore the probability that the residual mass contains only benign tissue. This will be reflected in the prediction model by a change in the regression coefficient for AFP. We may expect that a high serum level of AFP will still be associated with yolk sac elements in the primary tumour, but yolk sac will be less predictive for residual tumour. Although the association between AFP serum level and the presence of yolk sac elements in the primary tumour will remain the same, the validity of the model can hence change over time or between centres with different chemotherapy regimens.

The associations with the residual mass histology were expressed as odds ratios as a result of the logistic regression analysis. Instead of logistic regression analysis for diagnostic problems, Cox regression analysis may be used for prognostic problems. Analogously, the regression coefficients in Cox regression models can represent biological associations that are independent of time or place. Validation studies on a Cox model may therefore show similar hazard ratios as estimated in the development data. Instead of an intercept, a baseline survival function is fitted that can differ between populations. Discrimination is similarly affected as in logistic regression analysis. The model will discriminate less in more homogeneous populations. However, estimation of discrimination is less straightforward for prediction models, because censoring complicates the counting of observed outcomes.

In summary, the regression coefficients of the model for residual mass histology were similar in different samples. We may conclude that the modelled associations are probably generalisable to other populations. In contrast, the intercept of the model may change per setting. Therefore, it would be sensible to check the model intercept with own data before the model is adopted in a particular centre. If necessary, the intercept may be adjusted. Further research is required to develop sensible adjustment strategies. One may think of an empirical Bayes method.⁶ The model intercept estimated in the development data may serve as the prior. Taking this information into account, the adjusted model intercept for the new data will be in-between the original intercept and the intercept as estimated in the new data.

An intercept check or a more extensive validation study can only be performed, if relevant data are available. Particularly in retrospectively collected data, values can be missing. The proportion of missing values will influence the validation procedure. If few data are missing, the values may be imputed using the correlation with the other variables. If many data are missing from one single variable, validation of an alternative model without that variable may be considered. This is only possible if the development data are available to construct the alternative model. The consequences of validating a model in incomplete data is subject of further research. Also the proportion of missing values allowed in validation studies is debatable.

An update of the prediction model for residual mass histology

Once new data have become available, such as the validation data described in this thesis, updating of the prediction model may be considered. A model derived from more recent data and from a broader spectrum of centres will probably be better generalisable to new patients. In chapter 8, an updated prediction model was estimated based on twice as much data as the original model (1094 instead of 544 patients). The predictor effects of the updated model were slightly changed, which may improve the model performance for future patients.

Model updating intends to be universal, in contrast to local model adjustment in case of poor validity (Figure 9.3). If no differences in intercept and regression coefficients can be found between centres or in time, the updated model can be based on all available data. On the other hand, if differences are found, the data might be weighed differentially. More recent information can for instance be more relevant for future predictions than information from longer ago. Weighing strategies are not available yet for this type of analysis, and will probably be subjective.

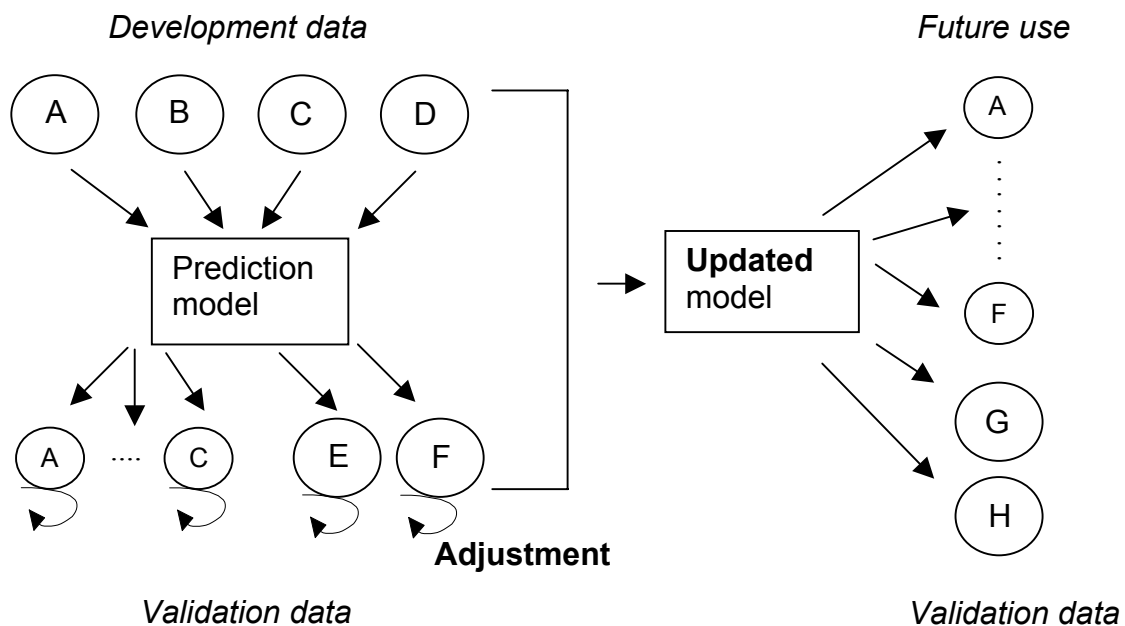


Figure 9.3 Schematic representation of model adjustment to improve the local validity versus model updating to improve the transportability. The transportability of the updated model to the centres G and H may be better than the transportability of the original model to E and F, since the updated model is based on more data and more centres.

Problems with implementation of clinical prediction models

Prediction models are built with empirical data and can therefore provide evidence-based predictions. Although the current opinion is to stimulate evidence-based medicine, many physicians prefer informal and non-quantitative predictions that are estimated by combining clinical experience and published evidence. I mention three reasons for this.

First, prediction models may not be valid. The models are developed in a limited number of patient data and the validity of the model in new patients needs to be assessed. Physicians are particularly worried about the correctness of the predicted probabilities for their own patients. Calibration plots with some deviations of the ideal line as shown in the chapters 5-7 may not be convincing enough. Further, the precision of the predicted probabilities may be questioned, especially if the development dataset was small which resulted in large confidence intervals.

Second, the use of prediction models is not always straightforward, since the models are developed to be statistically optimal and the associations between predictors and outcome can be quite complex. In the model for residual mass histology for instance, the square root and natural logarithm transformations were used. A user-friendly presentation is necessary. Therefore, a score chart was presented to facilitate a quick estimation of the predicted probability (see chapter 5). With this chart, a sumscore can easily be calculated. The sumscore corresponds to the predicted probability, which can be read from a figure. Also, a spreadsheet containing the exact formulas may be helpful (Figure 9.4). The physician needs only to fill in the predictor values and the spreadsheet shows immediately the predicted probability with the corresponding 95% confidence interval. An even more elegant approach would be to program the calculation of the prediction model in the electronic medical record system. Such a system could provide the predicted probability along with other test results.

Third, the models provide predicted probabilities, not merely a yes/no answer. For the interpretation of a probability a sensible threshold value is required. A threshold value is usually not explicitly determined for a particular problem. Nevertheless, common practice uses implicit threshold values. For instance, the European policy to resect all masses of 10 mm or larger is common practice, which corresponds with a threshold value of approximately 62% in the development data. The more stringent Norwegian policy of resection of all masses, except a specific subgroup had a threshold value of 85%. Therefore, a threshold value in between 62% and 85%, e.g. 70% or 80%, may be recommended for the model predicting residual mass histology. Alternatively, the analyst may already decide for the physician which threshold value should be used, instead of a recommended threshold value. In that case, the model can give a simple yes/no answer.

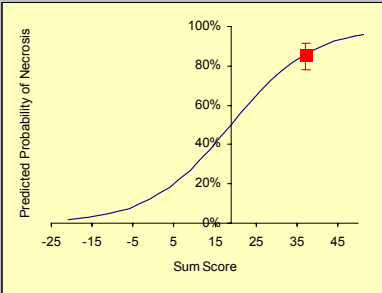
Conclusions

- At least three aspects of model performance are important for the assessment of validity of clinical prediction models: calibration; discrimination; and clinical usefulness. Ideally, all three aspects are perfect.
- Relevant measures are available to study each aspect of model performance. The power to detect a change in model performance in a new group of patients depends on the extent of the change and on the effective sample size. The effective sample size usually equals the smallest of the number of events and non-events rather than the total number of subjects.
- Current prediction models cannot predict risks of occult metastasis in clinical stage I patients above 50%. Selection of more predictors will probably improve the prediction models, which will likely result in higher predicted risks for specific patients.
- The previous developed prediction model for residual mass histology can be used to improve the selection of patients for resection. Carefulness is however required, because a prevalence correction (expressed by the model intercept) may be necessary.
- When new patient data have become available, more recent or from other centres, re-estimation of the regression coefficients is sensible. A model based on data from more centres will provide more accurate and more generalisable predictions.

Practical recommendations for external validation studies

- Adequate data should be available to study model validity. Coding and definitions of predictor variables should be the same as in the development set and the proportion of missing data should be low;
- Samples with a substantial size, e.g. at least 100 events and 100 non-events, should be used to obtain adequate power to detect poor validity;
- A validation plot with the observed frequencies against the predicted probabilities should be used. The plot can give insight in all relevant aspects of model performance, i.e. calibration; discrimination; and clinical usefulness;
- Particularly the model intercept, i.e. the outcome prevalence given the case-mix, should be studied before a model can be adopted in another setting.

Prediction model for residual mass histology in nonseminomatous testicular germ cell cancer			
Patient Nr:		14	
Normal value for LDH in local units		480	
Predictors		Value	Score
Primary teratoma-negative? yes=1 / no=0		1	11
Pretherapy level of AFP normal=1 / elevated=0		1	11
Pretherapy level of HCG normal=1 / elevated=0		1	7
Pretherapy level of LDH in local units		625	2
Pretherapy mass size in mm, minimal 2 mm		60	
Posttherapy mass size in mm, minimal 2 mm		20	-3
Change in mass size automatically calculated		67%	9
		Sumscore	37
<i>Predicted probability of benign tissue:</i>		85%	
<i>Confidence interval:</i>		77% - 90%	



<<http://www.eur.nl/fgg/mgz/software.html>>

Figure 9.4 Example of a spreadsheet to facilitate the probability estimation. The values of the six variables filled in for a particular patient will automatically give the total sumscore plus corresponding predicted probability. The 95% confidence interval can be read from the figure. Also the exact results from the formula are given.

References

1. Altman DG, Royston P: What do we mean by validating a prognostic model? *Stat Med* 19:453-473, 2000
2. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. *Ann Intern Med* 130:515-524, 1999
3. Habbema JD, Hilden J: The measurement of performance in probabilistic diagnosis. IV. Utility considerations in therapeutics and prognostics. *Methods Inf Med* 20:80-96, 1981
4. Braitman LE, Davidoff F: Predicting clinical states in individual patients. *Ann Intern Med* 125:406-412, 1996
5. Little RA: Regression with missing X's; a review. *J Am Stat Assoc* 87:1227-1237, 1992
6. Greenland S: Putting background information about relative risks into conjugate prior distributions. *Biometrics* 57:663-70, 2001

