Summary

Clinical prediction models combine patient characteristics to predict the probability of having a certain disease (diagnosis) or the probability that a particular disease state will occur (prognosis). The predicted probability of the diagnostic or prognostic outcome may assist the clinician in decision making for patient care. Before a prediction model can reliably be applied in clinical practice, the performance of the model in new patients ('external validity') needs to be studied. This thesis described several theoretical and practical aspects of the external validation of clinical prediction models. The objectives were i) to describe aspects of model validity and relevant performance measures; ii) to estimate the power of these performance measures; iii) to externally validate a prediction model for residual mass histology in testicular cancer; and iv) to update this model with all available information.

In chapter 2, three aspects of model performance are discussed: calibration; discrimination; and clinical usefulness. Calibration refers to the agreement between the observed outcome frequencies (e.g. the proportion of patients with a benign residual mass as opposed to residual tumour) and the predicted probabilities that the patients have the outcome (e.g. the predicted probability that the residual mass contains only benign tissue). For instance, if a 70% probability is predicted that the residual mass contains only benign tissue, then approximately 70 out of 100 patients with predicted probabilities of 70% should have benign tissue; the other 30 patients should have residual tumour. Discrimination refers to the ability of the model to distinguish a patient with benign tissue from a patient with residual tumour. Ideally, patients with residual tumour have predicted probabilities of benign tissue close to 0%, while patients with benign tissue have predictions close to 100%. Clinical usefulness indicates whether the model can be helpful to a clinician in making a correct decision, such as offering surgery to the patient.

Several measures were described for each performance aspect. Calibration is often studied with the Hosmer-Lemeshow statistic. This statistic compares the observed frequencies with the predicted probabilities per decile of predicted risks. A plot with the observed outcome frequencies against the predicted probabilities (validation plot) is more attractive, since it gives a visual impression of the calibration. The line in this plot can be described with an intercept and a slope. If the calibration is perfect, the intercept=0 and the slope=1. The values of these parameters can be statistically tested (separately or jointly). Discrimination is usually studied with the concordance (*c*)-statistic. For binary outcomes this statistic is identical to the area under the receiver operating characteristic curve, which is often used in the evaluation of diagnostic tests. The *c*-statistic may be interpreted as the

likelihood that a patient with the outcome (e.g. benign tissue) has a higher predicted probability for having the outcome, than a patient without the outcome (e.g. residual tumour) for a random pair of patients with different outcome values. Sensible measures for clinical usefulness are more difficult to define. In order to use a predicted probability for medical decision making, a threshold value is required. Decisions for patients with predicted probabilities above the threshold will be different from the decisions for patients with predicted probabilities below the threshold value. Given a threshold value, measures such as accuracy, sensitivity, and specificity can be calculated. However, more appropriate measures for clinical usefulness include also the consequences of the decisions, such as weighed error rate.

The power of several performance measures to detect changes in model performance was studied with power calculations and Monte Carlo simulations in chapter 3. A statistic that tests jointly the intercept and the slope of the calibration line of being respectively 0 and 1 had more power than the Hosmer-Lemeshow statistic to detect changes in calibration. The power of the *c*-statistic to detect changes in discrimination depended highly on the outcome prevalence. This result showed that the focus in power calculations should also be on the smallest of the number of events and non-events (the effective sample size), rather than only on the total sample size. As a rule of thumb, a minimum of 100 events and 100 non-events was suggested to obtain adequate power in external validation studies.

The external validity of a model does not only depend on the new patients for whom the model is applied, but also on the development process of the model. Therefore, this thesis contained also some illustrations of model development aspects. Chapter 4 reviewed the literature on predictors for the presence of occult metastasis in clinical stage I nonseminomatous testicular germ cell cancer. The strength of the predictors was estimated in a meta-analysis. The results of this analysis may guide a better selection of important predictors to include in a prediction model for occult metastasis. Such a prediction model will probably improve the predicted risks of occult metastases as estimated with present multivariable models. As a result, adjuvant therapy can be better targeted to those with likely occult metastasis, while surveillance can be more precisely offered to patients probably already cured.

The development process of the prediction model for residual mass histology in nonseminomatous testicular germ cell cancer was described in chapter 5. The model predicts the probability that a residual retroperitoneal mass contains only benign tissue after cis-platin based chemotherapy for metastatic tumour. Model development included variable selection with literature data (in total 996 patients), a detailed assessment of the form of the association of the predictors in individual patient data from six international study groups (n=544, development data), and correction for the introduced optimism in the estimated regression coefficients by a shrinkage factor. The final model contained six predictors. By

definition, a prediction model is on average well-calibrated in the data used for the development process. The model showed also good discrimination in the development data set (*c*-statistic=0.83, corrected for optimism). The external validation of the prediction model was good in a first validation set (n=172). The model showed adequate calibration and good discrimination (*c*-statistic=0.82).

Two other external validation studies were described in the chapters 6 and 7. Patient data were used from Indiana University Medical Center (IUMC, n=276) and an EORTC/MRC trial (n=105) The prevalence of benign tissue was lower in these two data sets (28% and 26% respectively) than in the development set (45%). In the IUMC data, many values of one predictor were missing and two other predictors were documented categorically, while the model uses continuous values. Therefore, an alternative model was constructed. The low prevalence of benign histology could partly be explained by the more severe patient characteristics, such as larger residual masses. The remaining, statistical non-significant, difference in prevalence had to be attributed to differences in patient characteristics which were not included in the model, such as referral pattern. The alternative model had reasonable discriminative ability (*c*-statistic=0.79).

The low prevalence of benign tissue in the EORTC/MRC trial (26%) was remarkable, since all patients had good prognosis disease. However, the patients of the trial selected for resection were only a small part of all good prognosis patients. Many good prognosis patients have small, often benign, residual masses, that are not resected. Therefore, the patients who are resected have relatively often residual tumour. The low outcome prevalence was completely explained by the variables in the model, i.e. the overall frequency of benign tissue was similar to the average predicted probability of benign histology. Low predicted probabilities were however in disagreement with the observed frequencies, while higher predicted probabilities were better calibrated. Since a physician will choose observation over resection only if the predicted probability for benign tissue is relatively high, the model could still be valuable in that decision making process. However, the distribution of the predicted probabilities showed that most predicted probabilities were under 70%, which may be considered as a sensible threshold value in this clinical problem. Many decisions based on the prediction model would have been the same as the actually made decisions, i.e. surgery. The model has therefore, little additional value for the studied patients. In contrast, a substantial part of the unresected patients had also predicted probabilities of benign tissue below 70%. For this group, a decision based on the model would have been different from the actually decision.

All available individual patient data of the development study and the validation studies were combined in chapter 8 (n=1094) to update the prediction model for residual mass histology. The regression coefficients of the updated model were only slightly different from the regression coefficients of the original model, but more precise due to the larger sample size. The updated model showed good performance in four of the six patient groups. Low predictions for the patients from the EORTC/MRC trial again disagreed with the observed

outcome frequencies. Predicted probabilities were systematically too high for patients from a specialised US centre. An adjustment of the model intercept may be considered for these patients.

This thesis concluded with a discussion on the topics addressed in earlier chapters, based on the formulated objectives (chapter 9). Further, practical recommendations for external validation studies were presented. With respect to study design, it was recommended to collect adequate data. Predictor definitions should be the same as in the development data and missing values should be limited. A substantial number of patients with the event and a substantial number of patients without the event are required to achieve adequate power to detect possibly poor validity. An essential element in the data analysis is a validation plot showing the observed frequencies against the predicted probabilities. The plot may give readily insight in all the relevant aspects of model performance.