Chapter 1

Introduction

Clinical prediction models are increasingly important, as apparent from the large number of recent publications that describe the development of such models. Publications on the validity of prediction models are less frequent, though recognised as of paramount importance.[1,2] The models intend to assist clinical decision making. Prediction models combine patient characteristics to estimate the probability of having a certain disease (diagnosis) or the probability that a particular disease state will occur in time (prognosis). Prediction models have been developed in a variety of clinical areas. Examples include internal medicine, orthopaedics, and fertility (Table 1.1).

Clinical prediction models should be developed with high-quality patient data. Since the predictions are commonly less accurate for new patients than for the patients used in the modelling process, assessment of the model validity in independent data is required (external validation).

Table 1.1 Some examples of clinical prediction models

Clinical area	*Prediction model*	*Outcome*	*Decision*
Internal medicine	APACHE and SAPS score[3,4]	Severe illness	Type of care (intensive care?)
Neurology	Glasgow coma scale[5]	Death	Type of care (intensive care?)
Orthopaedics	Ottawa ankle rule[6]	Fracture	Further diagnostic work up (radiography?)
Infectious disease	Meningitis model[7]	Bacterial meningitis	Type of treatment (antibiotics?)
Cardiology	GUSTO-I model[8]	30-day survival after acute myocardial infarction	Type of treatment (aggressive?)
Fertility	Eimers model[9]	Spontaneous pregnancy	Start treatment
Traumatology	TRISS and ASCOT[10,11]	Death	Evaluation of care (differences between centres?)

APACHE: acute physiology and chronic health evaluation
SAPS: simplified acute physiology score
TRISS: trauma score and the injury severity score
ASCOT: a severity characterisation of trauma

Model development

Most prediction models are developed with regression analysis techniques, although other techniques such as classification trees and neural networks are sometimes used. A diagnostic outcome is often modelled with logistic regression analysis and a prognostic outcome with Cox proportional hazard analysis, which models the time aspect of the prognostic outcome.

Several steps need to be taken, before the model parameters can be estimated.[12] First, choices need to be made on the coding of categorical predictor variables and the shape of the associations between continuous predictor variables and the outcome.[13,14] Often a continuous variable is included as a linear term, although the shape of the association with the outcome is not necessarily linear. For instance, U-shaped and J-shaped associations are also found in medicine.

Second, adequate predictor variables need to be selected. Usually, many candidate predictor variables are available of which a subset is selected. A common selection strategy is the stepwise selection procedure. This is an automated statistical procedure that considers step by step the additional predictive value of the variables. A predictor is selected if the association with the outcome is statistically significant in addition to the already included predictor variables. Statistical significance is indicated by a low p-value, such as $p<0.05$. However, naive use of a low p-value in small samples causes many statistical problems, including overestimation of the regression coefficients, and underestimation of standard errors.[12,15] Therefore, higher p-values for inclusion have been advocated, e.g. $p<0.20$ or $p<0.50$.[16]

Third, model assumptions can be assessed, such as the additivity of the predictor effects. Additivity implies that the effect of a predictor is expected not to modify the effect of another predictor in the model. For instance, the effect of the presence of a metastatic lesion on the 5 year-survival probability can be modelled the same in young patients and in old patients. If this effect is age-dependent however, the model should include an interaction term, which is the cross-product of the two predictors 'presence of metastatic lesion' and 'age'. The number of possible interaction terms is usually large. If 6 variables are entered in the model, already 15 first order interaction terms are possible. Therefore, a conservative attitude towards testing of interaction terms has been proposed.[17]

Model performance

Once a model is developed, the performance of the model needs to be studied. Two aspects are often considered: calibration and discrimination[20]. Calibration or reliability concerns the agreement between the outcome frequencies as observed in the data and the predicted probabilities of the model. If a model is developed to predict the probability that a residual mass after chemotherapy treatment for advanced testicular cancer is completely benign, the observed frequencies of benign tissue should agree with the predicted probabilities for benign tissue. Thus, if 100 patients have a predicted probability of 70% to have benign tissue, on average 70 out of the 100 patients should actually have benign tissue and 30

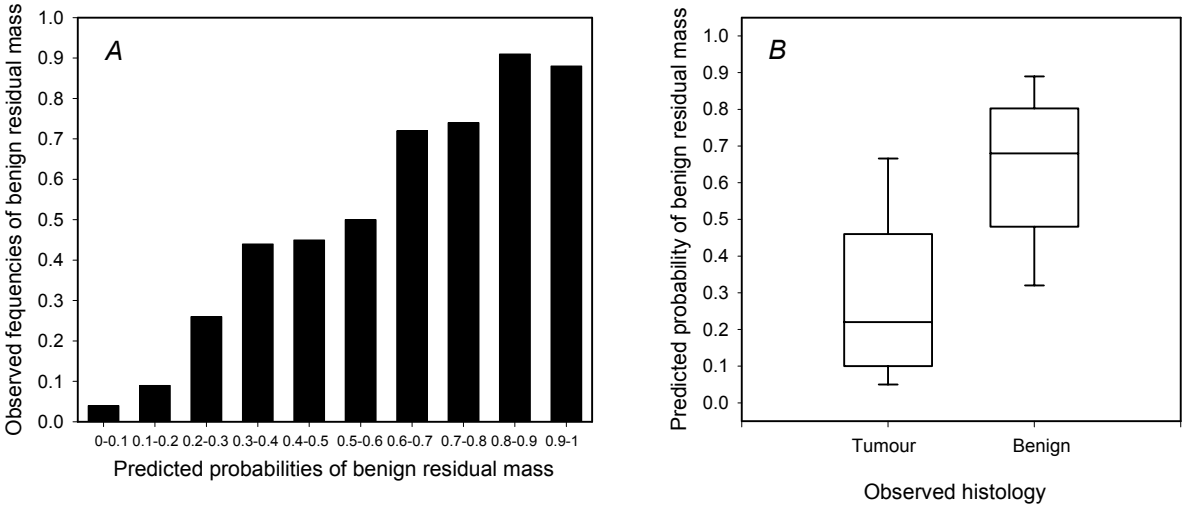

Figure 1.1 Calibration, i.e. agreement between observed frequencies and predicted probabilities (Figure A) and discrimination of predictions for residual mass histology (Figure B). Observed frequencies correspond with the predicted probabilities over the whole range of predicted probabilities (A). Figure B shows the distribution of the predicted probabilities by the observed histologies (tumour or benign). The boxes indicate the 25 and 75 percentile of predicted probabilities, the whiskers the 10 and 90 percentiles. Most predicted probabilities for masses containing tumour were under 50%, and most predicted probabilities for masses with benign tissue over 50%. Data were derived from 544 patients with testicular germ cell cancer.[21]

should have residual tumour. The observed frequencies and predicted probabilities may be compared for instance per 10% predicted probability intervals (Figure 1.1A). If the observed frequencies and predicted probabilities are in agreement over the whole range of predicted probabilities, the calibration of the model is good.

Discrimination refers to the ability of the model to distinguish a patient with the outcome from a patient without the outcome. A model that predicts for all patients the overall outcome frequency is well calibrated, but has no discriminative ability. A good discriminating model will predict high probabilities for patients with the outcome and low probabilities for patients without the outcome. In the example of residual mass histology, the model discriminates well if the patients with residual tumour have low predicted probabilities of benign tissue and the patients with benign tissue have high predicted probabilities (Figure 1.1B).

Apparent validation

The performance of the prediction model is usually tested in the dataset that was used for model development (development data). This apparent validity may give a first impression of the model performance. However, estimates of apparent validity will always be optimistic, since estimation (i.e. selection of variables, estimation of regression coefficients) and testing are performed in the same data.

Note that ‘validity’ is used in this thesis to indicate mainly the performance of the prediction model. The term ‘validity’ may also refer to the correctness of the selected

variables, the chosen variable transformations and the estimated regression coefficients. Here, these issues are considered to influence the performance and therefore the validity of a prediction model.

Internal validation

More reliable estimates of model performance can be derived by studying the internal validity (Table 1.2). This type of validity can be studied with cross-validation or resampling techniques, such as bootstrapping.[22] In the bootstrap procedure, many samples with the same size as the development data are drawn with replacement from the development data. In each bootstrap sample the development process is repeated. The resulting model is then tested in the original sample. If calibration and discrimination of the models derived from the bootstrap samples are adequate in the original sample, the internal validity is good.[23]

A common finding in internal validation studies is too extreme predicted probabilities, i.e. low predictions are too low and high predictions are too high. This is a result of overestimated regression coefficients ('overfitting'). Overfitting will particularly occur when the dataset is small, when many predictors are considered and when the dataset is used for all modelling steps.[24] The estimates of the regression coefficients can be corrected by multiplication with a so-called 'shrinkage factor', having a value between 0 and 1. The factor can be derived from the bootstrap samples.[12,25,26]

External validation

Before a prediction model can be widely used with confidence, the external validity needs to be assessed. External validity relates to the model performance in samples of patients from another, though related population.[1] For instance, similar clinical in- and exclusion criteria may have been used to define the sample, but the patients were treated more recently or in another centre. Time and place aspects may both influence the external validity of a prediction model. More recently treated patients may have received a new therapy that is more effective. This new therapy may then improve the prognosis of the patients. A disease may also be earlier diagnosed, e.g. as a result of an introduced screening program. In this case, the patient will be treated in an earlier disease stage, which may also improve the prognosis. Further, other medical centres may treat patients with different characteristics ('case-mix'). If the model is developed for example in patients aged between 20 and 40, application of the model to patients of 60 years of age requires extrapolation. An internally valid prediction model will not automatically be valid for these new patients.

Table 1.2 Hierarchy in the validation of clinical prediction models

Type of validity	*Required data*
Apparent	Sample used to develop the model (development data)
Internal	Resamples of development data derived with bootstrapping
External	Sample of new patients, treated either more recently or in another centre

To assess the performance of the model in patients treated more recently or in another centre, data from these patients are required. The data can be retrospectively collected. However, the data should be accurate and similar to the development data.[27] The number of missing values should be limited and definitions of variables should be the same.

Presentation

The formula of the final prediction model can be quite complex. The formula can include predictor transformations (e.g. square roots or logarithms) and non-linearities. For diagnostic outcomes modelled with logistic regression, for instance, the linear predictor (i.e. the sum product of the regression coefficients and the predictor values) corresponds to the natural logarithm of the odds of having the outcome. A model presentation in a user-friendly format, such as a score chart, may facilitate the probability estimation. A score chart assigns each predictor value a rounded score. The total score of the patient corresponds to the predicted probability. Prediction models can also be presented as a nomogram[18,19] or in a spreadsheet.

Adjusting and updating clinical prediction models

An external validation study may show that the predicted probabilities are poorly calibrated for patients from a new centre. The predictions can for instance be systematically too high. The average observed outcome frequency is in that case lower than the average predicted probability. To achieve model validity, adjusting the prediction model for this particular centre can be considered ('recalibration').[28]

Further, a prediction model can be updated, when a substantial amount of new data have become available, either from new centres or from more recently treated patients from the same centres. An updated model that is based on patient data from extra centres will probably be better transportable to a new centre than the original model, that was based on fewer centres. In addition, a model based on more recently treated patients may be more valid for newly treated patients than the original model.

Clinical applications in testicular germ cell cancer

This thesis describes model development and model validation aspects for two clinical decision problems in nonseminomatous testicular germ cell cancer (TGCC). The problems are similar in the sense that both depend on the uncertainty of the presence of tumour.[29]

TGCC is a rare disease and represents only 1% of all cancers in men. However, it is the most common solid malignancy in men aged between 20 and 34 years. The annual incidence of TGCC for this age group is approximately 4 per 100,000 white men and 0.9 per 100,000 black men. Germinal tumours result from abnormal development of embryonic pluripotent stem cells, which normally develop into spermatocytes producing cells. The tumours are divided into seminomas and nonseminomas reflecting the origin of the tumour and the ability to differentiate. A tumour is diagnosed as a seminoma, if the tumour contains pure seminomatous tissue and the serum level of alpha-fetoprotein (AFP), a marker of nonseminomatous tumour, is normal. Nonseminomas are composed of embryonal

carcinoma, teratoma, choriocarcinoma, yolk sac, or seminoma cell types. Nonseminomas may produce AFP, human chorionic gonadotropin (HCG), or lactate dehydrogenase (LDH) depending on the composition of the tumour.

Patients with abnormal findings on ultrasonography or elevated serum levels of AFP, HCG or LDH are diagnosed with TGCC. Usually, they first undergo orchidectomy to remove the primary tumour. Further, clinical staging procedures may reveal whether the tumour has spread through the body. The clinical stage of the disease determines the prognosis and further treatment of the patient. If serum tumour markers have been normalised after orchidectomy and radiological examinations of the abdomen and chest do not show further abnormalities, the disease is assumed to be limited to the testis, epididymis, or spermatic cord and is assigned 'clinical stage I'. Advanced disease includes metastases in the retroperitoneal lymph nodes (stage II), in the supradiaphragmatic lymph nodes (stage III), and extralymphatic metastases, e.g. in the lung, the liver, or in the bones (stage IV).

Occult metastases in clinical stage I testicular cancer

Around 30% of the patients with clinical stage I nonseminoma have occult metastases in the abdomen, which may become apparent at retroperitoneal lymph node dissection (RPLND) or during surveillance. The decision problem is whether the patient should be treated immediately after orchidectomy with either RPLND or chemotherapy or should be closely surveyed, with additional treatment if the patient does relapse. Immediate treatment is associated with risks of mortality and morbidity in case of RPLND or long term toxic effects in case of chemotherapy. Conversely, surveillance can lead to detecting relapses at a more advanced disease stage, particularly if the compliance for surveillance is poor, which may jeopardise the prognosis.

An important aspect of this decision problem is the risk of a patient to have occult metastasis. Patients at high risk for occult metastasis may be offered immediate adjuvant treatment, while such treatment may be unnecessary in patients at low risk. A systematic review was performed to quantify the strength of important predictors for occult metastasis (chapter 4). Further, an overview is given of the available multivariable models to predict the risks and to divide the patient group into low, intermediate, and high risk patients. Based on the multivariable models, risk-adapted treatment strategies have been introduced. According to these strategies, low risk patients go on surveillance, intermediate risk patients are offered RPLND, and high risk patients receive chemotherapy.

Residual mass histology in advanced testicular cancer

Patients treated with orchidectomy for a nonseminomatous germ cell tumour, who have signs of metastases such as elevated serum tumour markers or radiological abnormalities in the lymph nodes or in organs receive cis-platin based chemotherapy. The success of the chemotherapy can be monitored by the normalisation of serum tumour markers and the reduction in size of the metastases measured on CT. Patients with normalised markers but residual mass in the retroperitoneal lymph nodes can undergo RPLND to remove the mass

by surgery. However, many of the resected lymph nodes do not contain tumour. For these patients, surgery was unnecessarily performed.

Here, the decision problem is whether an individual patient should undergo RPLND with the risk of surgical morbidity and mortality. Alternatively, observation of a patient with residual tumour may worsen the prognosis by a higher risk of relapse, which is associated with a poor survival. Thus, only a high probability for benign tissue may justify observation. Therefore, a prediction model was developed to estimate the probability that the retroperitoneal lymph nodes contain only benign tissue after chemotherapy. This model was thoroughly validated in several datasets.

Rationale and outline of this thesis

Many clinical decisions may be supported by prediction models. The performance of the models needs to be assessed in external validation studies before the models can be applied with confidence in new patients. Although the need for external validation studies of prediction models is recognised, such studies are not routinely performed yet. A reason for this may be that guidelines for the study design are lacking. This thesis studies theoretical aspects of assessing the external validity of clinical prediction models. The focus is mainly on diagnostic outcome variables, modelled with logistic regression analysis. Further, the external validity of a model for patients with nonseminomatous testicular germ cell cancer is assessed.

The following specific objectives are defined:

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 patients.

Different aspects of model performance are described in chapter 2. Apart from calibration and discrimination, an aspect related to clinical usefulness, is also considered. An overview of relevant measures to study each aspect is given. The prediction model for residual mass histology developed with logistic regression analysis is used as an illustration. Furthermore, the validity of a survival model for prostate cancer is studied using the same performance measures. The power of the performance measures to detect poor validity is studied in chapter 3. This was done with power calculations and simulations.

The external validity of a model does not only depend on the new population in which it will be applied, but also on the development process of the model. Therefore, this thesis also contains two chapters on model development aspects (chapter 4 and chapter 5). Chapter 4 reviews the literature data on predictors for occult metastasis in clinical stage I nonseminomas. The quantified predictor effects may be used in the future development of a

model to predict the risk of occult metastasis. The development process of the prediction model for residual retroperitoneal mass histology in patients with nonseminomatous testicular germ cell cancer is described in chapter 5. The model predicts the probability that a residual retroperitoneal mass is completely benign. The model is externally validated in 172 patients from a similar population. Two other external validation studies using data of patients from Indiana University Medical Center (n=276) and an EORTC/MRC trial (n=105) are described in the chapters 6 and 7. All available data of the development study and the validation studies are combined in chapter 8 resulting in an updated model based on 1094 patients.

In the general discussion (chapter 9) the empirical results (chapters 4-8) are reviewed in the light of the theory described in the chapters 2 and 3. In addition, a measure for clinical usefulness is proposed, which elaborates further on measures described in chapter 2. The general discussion ends with some general recommendations for validation studies of clinical prediction models.

References

1. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Ann Intern Med 130:515-524, 1999
2. Altman DG, Royston P: What do we mean by validating a prognostic model? Stat Med 19:453-473, 2000
3. Knaus WA, Zimmerman JE, Wagner DP, et al: APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9:591-7, 1981
4. Le Gall JR, Loirat P, Alperovitch A: Simplified acute physiological score for intensive care patients. Lancet 2:741, 1983
5. Rimel RW, Jane JA, Edlich RF: An injury severity scale for comprehensive management of central nervous system trauma. Jacep 8:64-7, 1979
6. Stiell IG, Greenberg GH, McKnight RD, et al: A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med 21:384-90, 1992
7. Spanos A, Harrell FE, Durack DT: Differential diagnosis of acute meningitis. An analysis of the predictive value of initial observations. JAMA 262:2700-2707, 1989
8. Lee KL, Woodlief LH, Topol EJ, et al: Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. GUSTO-I Investigators. Circulation 91:1659-68, 1995
9. Eimers JM, te Velde ER, Gerritse R, et al: The prediction of the chance to conceive in subfertile couples. Fertil Steril 61:44-52, 1994
10. Boyd CR, Tolson MA, Copes WS: Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. J Trauma 27:370-8, 1987
11. Champion HR, Copes WS, Sacco WJ, et al: A new characterization of injury severity. J Trauma 30:539-45; discussion 545-6, 1990
12. Harrell FE, Lee KL, Mark DB: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy,and measuring and reducing errors. Stat Med 15:361-387, 1996
13. Harrell FE, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic prediction. Stat Med 3:143-152, 1984
14. Royston P, Ambler G, Sauerbrei W: The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 28:964-74, 1999
15. Steyerberg EW, Eijkemans MJ, Habbema JD: Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 52:935-42, 1999
16. Steyerberg EW, Eijkemans MJC, Harrell FE, et al: Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 19:1059-1079, 2000
17. Steyerberg EW, Eijkemans MJC, Harrel FE, et al: Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 21:45-56, 2001
18. Harrell FE, Margolis PA, Gove S, et al: Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological agents of Pneumonia,

Sepsis and Meningitis in Young Infants. WHO/ARI Young Infant Multicentre Study Group. Stat Med 17:909-44, 1998
19. Kattan MW, Wheeler TM, Scardino PT: Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. J Clin Oncol 17:1499-1507, 1999
20. Hilden J, Habbema JDF, Bjerregaard B: The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. Meth Inform Med 17:227-237, 1978
21. Steyerberg EW, Keizer HJ, Fosså SD, et al: Prediction of residual retroperitoneal mass histology following chemotherapy for metastatic nonseminomatous germ cell tumour: multivariate analysis of individual patient data from six study groups. J Clin Oncol 13:1177-1187, 1995
22. Efron B, Tibshirani RJ: An introduction to the bootstrap. New York, Chapman & Hall: London, 1993
23. Efron B, Tibshirani RJ: Improvements on cross-validation: the .632+ bootstrap method. JASA 92:548-560, 1997
24. Spiegelhalter DJ: Probabilistic prediction in patient management and clinical trials. Stat Med 5:421-433, 1986
25. Copas JB: Regression, prediction and shrinkage. JR Stat Soc B 45:311-354, 1983
26. van Houwelingen HC, Thorogood J: Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med 14:1999-2008, 1995
27. Braitman LE, Davidoff F: Predicting clinical states in individual patients. Ann Intern Med 125:406-412, 1996
28. van Houwelingen HC: Validation, calibration, revision and combination of prognostic survival models. Stat Med 19:3401-3415, 2000
29. Bosl GJ, Motzer RJ: Testicular germ-cell cancer. N Engl J Med 337:242-253, 1997