Chapter 3

Sample size considerations for external validation studies of prediction models: simulations with logistic regression

Abstract

The performance of a prediction model is usually worse in external validation data than in the development data. We studied at which effective sample sizes (i.e. number of events) relevant differences in model performance can be detected with adequate power. We used a logistic regression model to predict the probability that retroperitoneal lymph nodes of patients treated with chemotherapy for metastatic testicular germ cell cancer contained only benign tissue. The development data contained 544 patients of whom 245 had benign tissue at resection (45%). For three different sample sizes (n=100, 45 events; n=200, 90 events; and n=500, 225 events), the power of several performance measures was estimated with power calculations and Monte Carlo simulations. The power calculations and Monte Carlo simulations showed similar results. A model predicting probabilities that were on average 1.5 times too high on the odds scale could be detected as statistically significant different in 37%, 64% and 96% of the simulated validation samples, respectively. A decrease in the c-statistic from 0.83 to 0.73 was identified in 48%, 77%, and 97% of the validation samples. In conclusion, at least 90 events were required to detect substantial differences in model performance. We therefore suggest a minimum of 100 events and 100 non-events to obtain adequate power in validation studies.

Introduction

Predictive logistic regression models are important tools to provide estimates of patient outcome probabilities. A model accurately predicting probabilities for patients in the development data does not guarantee accurate predictions for new patients from related populations, e.g. patients treated more recently or patients from another centre. Therefore, the performance of prediction models needs to be tested in new patients (external validation).[1,2] A straightforward approach to study external validity is to split the development data into two parts: one part containing early treated patients to develop the model and another part containing the most recently treated patients to assess the performance. With this approach, the temporal aspect of external validity may be studied.[3,4] Similarly, the place aspect can be studied by splitting the data according to treatment centres.[5,6] In new data, both place and time aspects can be studied.[7]

Validation studies may typically show a systematic deviation of the probabilities or too extreme probabilities.[8] A systematic deviation of the probabilities (overall too high or too low), suggests that an important predictor variable was not included in the model.[9] If the predicted probabilities were too extreme (i.e. high predictions too high and low predictions too low), the regression coefficients of the prediction model were on average too large.[10,11] Individual regression coefficients can also be incorrect, due to differences in predictor definitions (bias) or imprecise estimates of the coefficients (imprecision). Further, a different distribution of predictor values ('case-mix') can influence some aspects of model performance.

Common measures to assess model performance include i) calibration measures which study the agreement between observed outcome frequencies and predicted probabilities, ii) discrimination measures which study the ability of the model to distinguish between patients with different outcomes, and iii) overall performance measures, which incorporate both aspects of calibration and discrimination.[12,13] Each measure has its own properties. A calibration measure will likely have more power to detect systematically deviating predictions than a change in case-mix, which is expected to affect mainly the discriminative ability.

Little is known about adequate sample sizes to study model performance in other populations.[8,14] Particularly, the use of too small samples may lead to statistically non-significant results, while true differences do exist. For binary outcomes, the power is also determined by the number of events, i.e. the effective sample size. For instance, a sample with 821 patients may seem large, but an outcome frequency of 1% implies that the sample contains only 9 events and may have therefore little power.[15] Here, we study at which number of events relevant differences in model performance can be detected with measures for calibration and discrimination. We used a model that predicts the histology of retroperitoneal lymph nodes in patients treated with chemotherapy for metastatic testicular germ cell cancer.[5,16] Evaluations were performed with power calculations and Monte Carlo simulations. We will show that relatively many events are required to obtain a reasonable power in validation studies.

Data and Methods

Prediction model for metastatic testicular germ cell cancer

Resected retroperitoneal lymph nodes of patients treated with chemotherapy for metastatic testicular germ cell cancer contain only benign tissue in about 45% of the operated patients. Those patients are unnecessarily operated. A logistic regression model was constructed to predict the probability of benign tissue. We used data of 544 patients of whom 245 (outcome frequency equals 45%) had benign tissue (development data set).[5] The model contained six predictor variables: three dichotomous variables (elevated prechemotherapy levels of the serum tumour markers alpha-fetoprotein [AFP] and human chorionic gonadotropin [HCG], and presence of teratoma elements in the primary tumour [ter]) and three continuous variables with transformation (natural logarithm of the standardised level of prechemo-

therapy lactate dehydrogenase [ln(LDH)], square root of the maximum mass size after chemotherapy [sqpost] and change in mass size during chemotherapy per 10 percent [change10]). The predicted probability of benign tissue ($\pi(X_i)$) was calculated by the logistic transformation: $\pi(X_i) = 1/(1+\exp[-\beta_0-X_i\beta])$, where X_i is a vector of the predictor values of patient i, β a vector of the regression coefficients and β_0 the intercept of the model. $\beta_0+X_i\beta$ is also known as the linear predictor (lp_i).

Simulation of differences in model performance

Random samples were created with Monte Carlo simulations by drawing covariate patterns from the development data set with replacement. For each covariate pattern X_i, lp_i was calculated. The outcome variable (Y_i) was generated by comparing the logistic probability $\pi(X_i)$ with an independently generated variable u_i having a uniform distribution from 0 to 1. We used the rule $Y_i = 1$ if $\pi(X_i) \geq u_i$ and $Y_i = 0$ otherwise. Sample sizes were 100, 200 or 500 with 45, 90, and 225 events respectively at an outcome frequency of 45%.

To simulate samples different from the development data, lp_i was changed into $lp.new_i$, which was used to generate $Y.new_i$. In the simulated validation samples the performance of the original model was then studied by comparing $Y.new_i$ with lp_i . This mimics the situation that the original model is tested in a validation data set, which is derived from another population where lp.new holds rather than lp. Hence, we are testing a model, which is incorrect for the simulated patients.

The following scenarios were simulated (Table 3.1). Predicted probabilities were systematically too high by subtracting a constant factor *a* from the linear predictor; lp.new = lp - *a*, with *a* = ln(1.5) (Scenario I) or *a* = ln(2) (Scenario II). This corresponds with predictions of the validation data set being 1.5 or 2 times too high on the odds scale. Too extreme predictions were simulated by multiplying the original linear predictor with *s*; lp.new = *s* * lp, with *s* = 0.8 (Scenario III) or 0.6 (Scenario IV). The multiplication reflects the common situation of overoptimism.[1,17]

To simulate a situation in which all coefficients were changed, we used the coefficients as estimated for another population (EORTC/MRC trial, Scenario V).[18] Further, a different case-mix was created using only a subsample of the development data, i.e. all patients with unfavourable values for at least the predictor variables ter, AFP, and change10 (Scenario VI). In this scenario, the outcome values were simulated with $lp.new_i = lp_i$. The scenario simulates a narrowed distribution of probabilities, which may affect the discriminative ability of the model. The intercept of the linear predictors used for the simulations were adjusted such that the average outcome frequency remained 45% for the Scenarios III and IV.

Table 3.1 Simulated clinical scenarios to study the power of performance measures for a prediction model in metastatic testicular germ cell cancer

Scenario	*No*	*Technique*	*Case-mix*
Predictions reliable	0	$lp_{new} = lp$	All patients
Predictions too high	I	$lp_{new} = lp - ln(1.5)$	All patients
	II	$lp_{new} = lp - ln(2)$	All patients
Predictions too extreme	III	$lp_{new} = 0.8*lp$	All patients
	IV	$lp_{new} = 0.6*lp$	All patients
Incorrect coefficients	V	$lp_{new} = X_i$*coefficients as estimated in an EORTC/MRC trial	All patients
Predictions reliable, but other case-mix	VI	$lp_{new} = lp$	Subsample of patients with at least three unfavourable predictor values (ter, AFP, change10) for benign tissue

lp_{new}: linear predictor to derive the outcome values for new patients of the validation samples
lp: linear predictor derived from the prediction model

Performance measures

The model performance was quantified with respect to calibration and discrimination. Calibration, or reliability, refers to the agreement between observed outcome frequencies and predicted probabilities. Discrimination refers to the ability to distinguish patients with different outcomes. Overall performance measures incorporate both calibration and discrimination aspects.

Calibration was studied with calibration curves, i.e. a graphical presentation of the relationship between the observed outcome frequencies and the predicted probabilities. Calibration curves can be approximated by a regression line (or calibration line) with intercept (α) and slope (β). These parameters can be estimated in a logistic regression model with the event as outcome and the linear predictor as only covariate.[19] Well-calibrated models have $\alpha = 0$ and $\beta = 1$. Therefore, a sensible measure of calibration is a likelihood ratio statistic testing the null hypothesis that $\alpha = 0$ and $\beta = 1$. The statistic has a χ^2-distribution with 2 degrees of freedom ('Unreliability' (*U*)-statistic).[1,13,19] The Hosmer-Lemeshow statistic for external validity was also estimated. To compute this statistic, predicted probabilities were divided in deciles. Per decile the observed frequency was compared with the average predicted probability. The statistic has a χ^2-distribution with 10 degrees of freedom.[20] Discrimination, i.e. whether the relative ranking of individual predictions is in the correct order, was quantified with the concordance (*c*)-statistic. This statistic is identical to the area under the receiver operating characteristic curve for binary outcomes.[11] The statistic indicates the proportion of all pairs of patients with different outcome values for which the patient with the outcome has a higher predicted probability

than the other patient. Measures used to quantify the overall performance were the Brier score[21], i.e. $\Sigma (Y_i - \pi_i)^2/n$ and Nagelkerke's R^2, i.e. a measure of explained variation calculated on the log-likelihood scale.[22]

The behaviour of the performance measures was studied for the different simulated scenarios. To obtain stable estimates, 100,000 patients were simulated.

Assessment of power

Since calibration is studied as the agreement between observed outcomes and predicted probabilities within the same sample, we used standard one-sample tests for the calibration measures (intercept, slope, *U*-statistic, and Hosmer-Lemeshow statistic). In contrast, the discriminative ability was compared between the development and validation settings and was therefore studied with two-sample tests.

The power to detect a statistically significant difference in model performance at a particular sample size was first calculated with standard formulas based on the Normal distribution. The number of events is implicitly incorporated in these formulas via the standard deviation of the difference in performance measure (σ). σ changes when the proportion of events (the outcome frequency) changes. Samples with lower outcome frequencies correspond to a larger σ. σ was estimated using the data of the development data set.

The formula to calculate the power given the outcome frequency is:

$Z_\beta = \sqrt{\frac{N\delta^2}{\sigma^2}} - Z_{½\alpha}$; with N = sample size, δ = difference in model performance, Z_β = value of the standard Normal distribution corresponding to β, with β=type II error rate and 1-β = the power of the hypothesis test, $Z_{½\alpha}$ = value of the standard Normal distribution corresponding to ½ α, with α = type I error rate (here: 0.05).

As an example, the power can be calculated to detect a miscalibrated model, that predicts on average 1.5 times too high probabilities ($\delta = \ln(1.5) = 0.4$, $\sigma = se/\sqrt{n} = 0.11/\sqrt{544} = 2.45$) with a sample of 100 patients as:

$Z_\beta = \sqrt{\frac{100[0.4]^2}{2.45^2}} - 1.96 = -0.30$ corresponding to $\beta = 0.62$ and $1 - \beta = 0.38$.

We calculated the power to detect the differences in intercept and slope of the calibration line, and the *c*-statistic for the three sample sizes (n=100, 200, 500). We also calculated the required sample size to achieve 80% power.

Second, we studied the power with the validation samples derived with the Monte Carlo simulations. To take the variability of the development data into account for the discrimination (two-sample t-tests), 5000 development samples of size 544 were drawn with replacement. For each sample the optimism-corrected *c*-statistic[1], and its standard error[23] were estimated. The *c*-statistic estimated per simulated validation sample could then be compared with a development sample using the two-sample t-test. The power of the *c*-statistic was determined by the proportion of p-values below 5% in 5000 simulated validation samples.

Table 3.2 Power and sample size calculations to detect a decrease in model performance for a prediction model in metastatic testicular germ cell cancer

Measure	*Original value*	*SE* [a]	*New value*	*Sce-nario*	Power			*Required sample size for 80% power (events)*	
					n=100	*n=200*	*n=500*		
Intercept|Slope=1	0	0.11	-ln(1.5)	I	38	65	96	281	(107)
			-ln(2.0)	II	81	98	100	99	(34)
Slope	1	0.09	0.8	III	16	27	56	882	(397)
			0.6	IV	47	76	99	221	(99)
c-statistic	0.83	0.02	0.73	V	59 [b]	82 [b]	97 [b]	174	(70)

[a] standard error, estimated in the development data (n=544)
[b] two-sample test

Results

Simulated scenarios

Figure 3.1 shows the calibration curves and performance measures of the prediction model corresponding to the simulated scenarios, at a very large sample size (n = 100,000). A sample from the same underlying population as the development data (Scenario 0) showed perfect calibration (slope = 1.0, intercept = 0.0), with *c*-statistic = 0.83, R^2 = 0.41, and Brier score = 0.17. Systematically too high predictions (Scenarios I/II) influenced the calibration by a change in intercept, but did not influence discrimination (*c*-statistic remained 0.83). The overall performance measures were poorer (R^2 decreased; Brier score increased slightly), because of the reduced calibration. Too extreme predictions (Scenarios III/IV) resulted in a decreased calibration slope and *c*-statistic. The overall performance measures reduced more markedly compared with systematic changes in the predictions (Scenarios I/II), because both calibration and discrimination were influenced. A scenario in which the correct model included regression coefficients as estimated in the EORTC/MRC trial (Scenario V), showed a reduction in all aspects of model performance. A decrease in heterogeneity influenced particularly the discriminative ability (Scenario VI; *c*-statistic = 0.73). A decrease in *c*-statistic from 0.83 to 0.73 may be interpreted as an increase of incorrect ranked pairs of patients with different outcome values from 17% to 27%, which is 1.6 times as high. In this scenario, the Brier score was closer to 0 (0.08), since the samples of Scenario VI contained less patients with benign tissue (10% versus 45% in Scenario 0). Thus the positive influence of the low average outcome frequency on the Brier score was larger than the negative influence of the reduction in discrimination. The average outcome frequencies in the Scenarios I-V were 38%, 34%, 45%, 45%, and 40%.

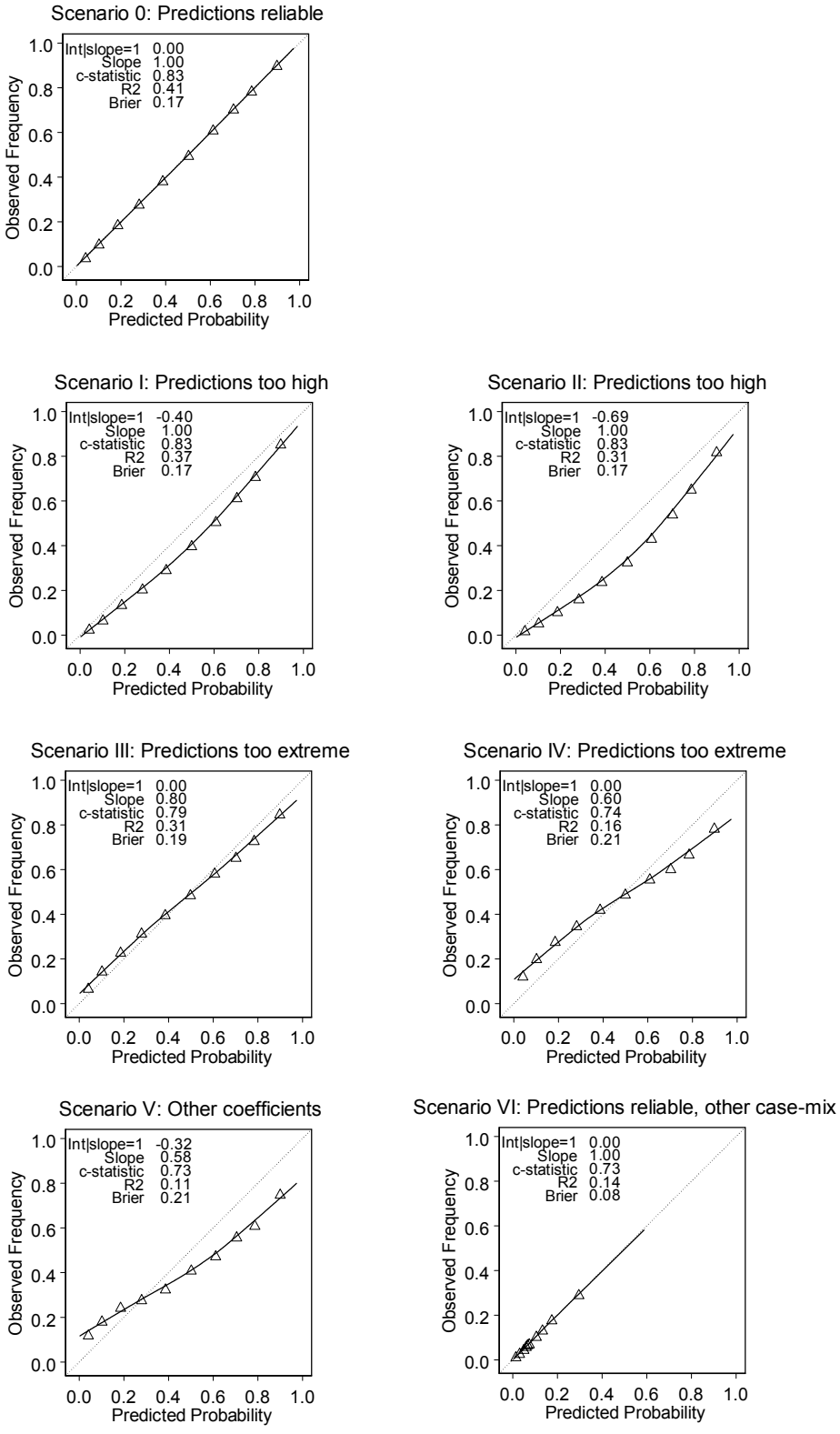

Figure 3.1 Calibration curves corresponding to different simulated scenarios (n = 100,000): development data (Scenario 0); systematically too high predicted probabilities (I/II); too extreme probabilities (III/IV); different regression coefficients (V/VI); different case-mix (VII). The dotted line indicates perfect calibration, i.e. observed frequencies and predicted probabilities are in complete agreement; the continuous line shows the relation between observed frequencies and predicted probabilities. Triangles indicate observed frequencies per decile of predicted probabilities.

Power and sample size calculations

Table 3.2 shows the power of the performance measures as calculated with standard formulas. Predictions that were twice too high on the odds scale could well be detected with the calibration-intercept in samples of size 100 (with 34 events, given the average outcome frequency of 34%, power = 81%). Typical overoptimism reflected by a slope of 0.8 may be detected in only 56% of all samples of size 500 (with 225 events). To achieve 80% power, a sample of size 882 (with 397 events) would be required. A change in the *c*-statistic from 0.83 to 0.73 at an average outcome frequency of 40% could be well detected in samples of 200 patients (with 80 events, power = 82%). At a an outcome frequency of 10%, a sample of 886 patients (with 89 events) would be required to achieve 80% power.

Power in simulated samples

Figure 3.2 shows the power of the calibration measures and the *c*-statistic to detect differences in model performance at 3 sample sizes (n = 100, 200, and 500). In general, the results are similar to the results of the power calculations in Table 3.2. The null hypothesis of no difference in model performance was true in samples of the same population as the development data (Scenario 0), and was rejected in around 5% of the samples.

The power to detect a model with on average 1.5 times too high predictions was 37%, 64%, and 96% (sample sizes of 100, 200, and 500 with 38, 76, and 190 events) in Scenario I and 80%, 98%, and 100% (34, 68, and 170 events) in Scenario II (predictions twice too high). The Hosmer-Lemeshow statistic had only 38% power to detect the lack of performance in samples of Scenario II with n = 100 (34 events) compared with 71% power for the *U*-statistic.

Typical overoptimism (slope = 0.8, Scenario III) was detected with the slope of the calibration line in 65% of all simulated samples of 500 patients (225 events). Samples of 100 patients (45 events) had poor power to detect overoptimism in Scenario III (power ≤ 20% for the different performance measures). The large overoptimism simulated in Scenario IV (slope = 0.6) was well detected with the slope, the *U*-statistic, and Hosmer-Lemeshow statistic in samples of 200 patients (90 events) or more. The proportion of samples in which a change in intercept was detected, should completely be attributed to the type I error. The proportions ranging from 6.3% - 9.5% indicate that the intercept, given that the slope was set to 1, is not normally distributed under the alternative hypothesis.

All measures had reasonable or good power to detect lack of performance if the regression coefficients were different (Scenario V). The power to detect the decrease in *c*-statistic in Scenario VI (from 0.83 to 0.73) was lower (13%, 26%, and 68%) than the power in the Scenarios IV and V (45%, 72%, and 95%; or 48%, 77%, and 97%), while the size of decrease in *c*-statistic was similar (around 0.1). This was caused by the lower average outcome frequency (10% versus 45% and 40%) resulting in different number of events (10, 20, and 50 versus 45, 90, and 225; and 40, 80, and 200, respectively). The power became comparable when we increased the average outcome frequency to 45% in Scenario VI.

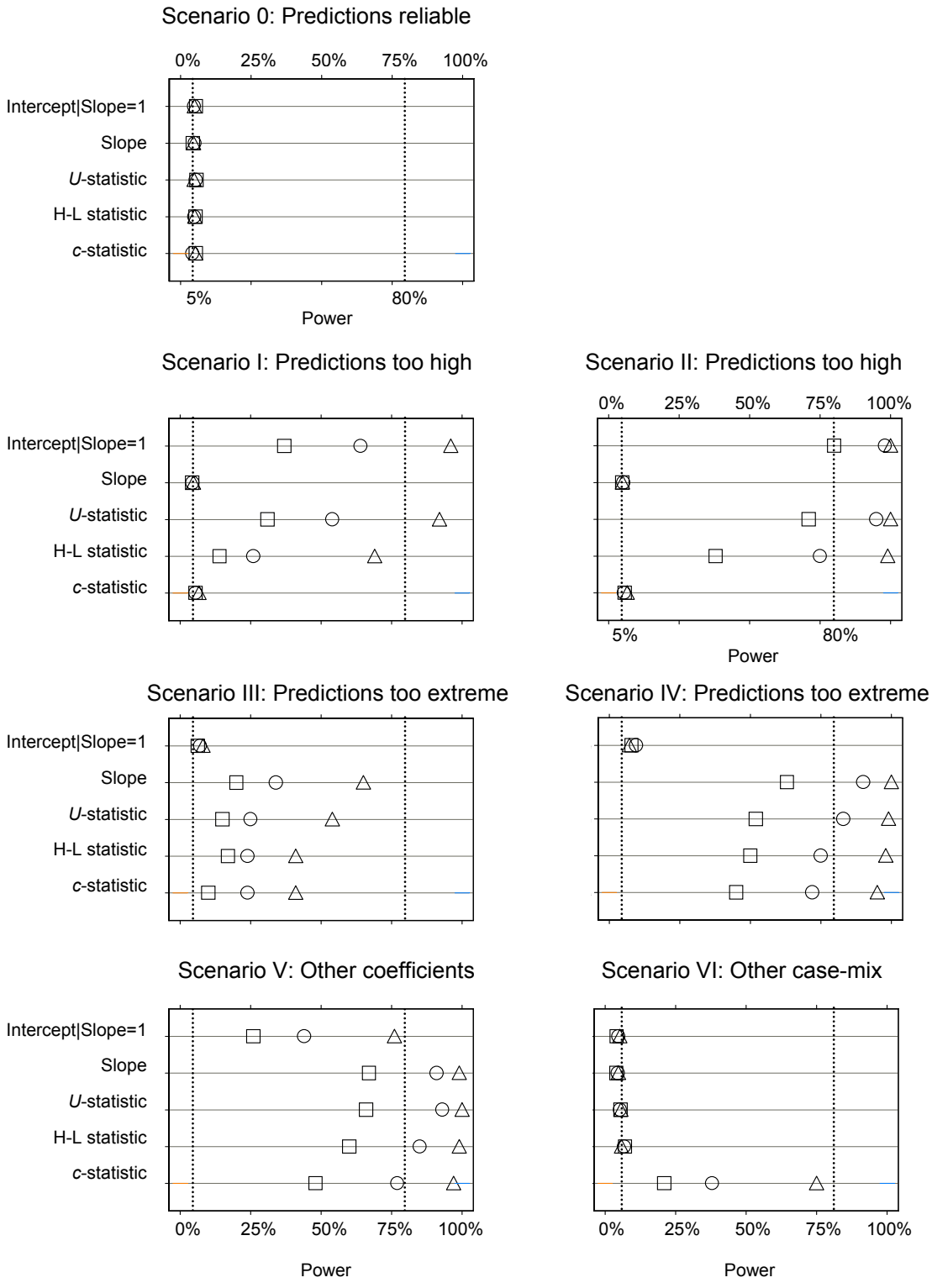

Figure 3.2 Power in percentages to detect differences in performance of a prediction model for several clinical scenarios at three sample sizes (square, n=100; circle, n=200; and triangle, n=500). The nominal Type I error was set at 0.05 (first dotted line). Power was considered good at 80% (second dotted line) or higher.

Discussion

We have shown that a substantial sample size and number of events are required to detect relevant decreases in model performance with a power of around 80%. A model with systematically too high predictions could be best detected with a test for a change in the intercept of a line describing the relation between the observed outcome frequencies and the predicted probabilities of the model (calibration line).[1] This measure was more specific than the Hosmer-Lemeshow goodness-of-fit statistic.[20] The *U*-statistic testing the intercept and slope of the regression line jointly had intermediate power. [1,13,19] The power to detect too extreme predictions (overoptimism) was highest for the calibration slope, followed by the *U*-statistic, Hosmer-Lemeshow statistic, and *c*-statistic, which all had similar power. A decrease in discriminative power expressed as a decrease in the *c*-statistic could be better detected in samples with more numbers of events, reflected by a higher average outcome frequency.

As expected, the more specific measures, such as intercept and slope of the calibration line, had the highest power to detect systematic deviations and overoptimism, respectively. The Hosmer-Lemeshow statistic had particularly little power to detect systematic deviations. This statistic has been introduced as a goodness-of fit-test for model development.[20] It has been shown that this statistic has reasonable power to detect whether non-linearity terms or alternative link functions are required.[24] Therefore, the Hosmer-Lemeshow statistic may better be reserved for model development and not for validation purposes. A good alternative in model validation may be the *U*-statistic, which tests jointly the intercept and calibration slope.

The power calculations and the Monte Carlo simulations showed similar results. This indicates that the assumed normality for the performance measures to calculate the power was correct, although the calibration intercept, given the slope was set to 1, showed some deviation under the alternative hypothesis. The power can hence easily be calculated, under the assumption that the 'virtual' standard deviation (σ) estimated in the development data equals the σ of the validation data. σ depends on the average outcome frequency. If σ can not reliably be estimated, simulation studies will be valuable. Furthermore, Monte Carlo simulations may be used to study the power of performance measures that do not follow a Normal distribution, such as the overall performance measures Brier score and R^2. To allow for the extended left hand tail of the 95% confidence intervals, a method different from the simple σ calculation is required, for example bootstrapping. This should then be applied to each simulation sample. Unfortunately, such a method was not feasible with our computer capacity.

The power of the performance measures was studied with a model containing six predictor variables with a *c*-statistic of 0.83 in the development data that had an average outcome frequency of 45%. The results should be interpreted in the light of this particular situation. Scenario VI showed the influence of the average outcome frequency on the power of the *c*-statistic. A decrease from 0.83 to 0.73 was detected in only 33% of the simulation samples containing 200 patients at an average outcome frequency of 10%. Simulated

samples with an average outcome frequency of 45% and the same heterogeneity as in Scenario VI had 88% power at sample size 200 to detect the decrease in the *c*-statistic (data not shown). Thus, studying a decrease in model performance for populations with much lower or higher average outcome frequencies will require larger samples. These findings confirm that the effective sample size is determined by the number of events (the average outcome frequency times the sample size) or non-events (in case of a frequent outcome), rather than the total number of subjects.[11,17,25] The number of events (or non-events) also sets limits to model development. Harrell and others proposed to use only one degree of freedom per ten events for model development.[11] A more unfavourable ratio might result in a poorly validating model.[1] In accordance with the 1:10 rule, we can formulate a rule of thumb for the minimum number of events and non-events required to properly study model performance. Here, samples of size 200 had approximately 80% power to detect substantial differences in model performance such as 1.5 times too high or low predicted odds, too extreme predictions expressed with a calibration slope of 0.6 or a decrease in the *c*-statistic of 0.1. At the average outcome frequency of 45%, 200 subjects corresponds to 90 events. At a 10% average outcome frequency, 88 events (n = 880) were required to detect a change in the *c*-statistic of 0.1. As a rule of thumb, we suggest a minimum of 100 events and 100 non-events to study model performance in a new population. For the detection of smaller differences in model performance even more events and non-events will be required.

In the literature, sample sizes of validation sets differ a lot.[26,27] If the model is developed in a training set and subsequently validated in a test set, the size of the test set is often relatively small.[28-30] Literature examples, which were used as illustration by Altman and Royston[8] described validation data with sizes varying between 52 and 479 patients. The number of events were much lower than 100 for most validation data (with a range from 24 to 115).[31-35]

There are several limitations to our study. First, the chosen clinical scenarios for which the power was studied, are arbitrary. Since we studied substantial changes, such as predicted probabilities of 1.5 times too high on the odds scale and too extreme predictions that should have been shrunk with a factor of 0.8 to 0.6, we consider the results informative. Other possible scenarios may be a misspecified continuous predictor effect, a missed interaction term, or an inappropriate transformation to link the linear predictor with the average outcome frequency corrected for case-mix.

Second, the validity of a prediction model in a new population depends also on the data used to develop the model. A large development sample will result in more confident estimates of the regression coefficients than a small sample. The likelihood that the model will perform well in new patients, particularly if the new patients are derived from a very similar population, may be higher for a large sample. However, we did not take this development uncertainty into account for the calibration measures. Our reasoning was that a literature model that is to be adopted by others, will be often considered as fixed. The uncertainty in the predicted probabilities is ignored then. In contrast, the uncertainty was

taken into account in studying discrimination, because discrimination is also determined by the study population.

In conclusion, we recommend to estimate separately the intercept and slope of the calibration line when validating the performance of a model from the literature. These measures had the highest power to detect miscalibration. The *U*-statistic testing intercept and slope jointly can be used as an overall calibration test. The *c*-statistic is useful to study differences in discrimination. We further suggest to use a sample with at least 100 events and 100 non-events. Many validation studies that did not show clear differences in performance might actually have been underpowered.

References

1. Harrell FE Jr, 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
2. Hand JD: Construction and assessment of classification rules. Chisester, England, John Wiley & Sons Ltd, 1997
3. Picard RR, Berk KN: Data splitting. Am Statistician 44:140-147, 1990
4. Miller ME, Langefeld CD, Tierney WM, et al: Validation of probabilistic predictions. Med Decis Making 13:49-58, 1993
5. 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
6. van Houwelingen HC, Thorogood J: Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med 14:1999-2008, 1995
7. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Ann Intern Med 130:515-524, 1999
8. Altman DG, Royston P: What do we mean by validating a prognostic model? Stat Med 19:453-473, 2000
9. Vergouwe Y, Steyerberg EW, Foster RS, et al: Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. J Urol 165:84-88, 2001
10. Copas JB: Regression, prediction and shrinkage. JR Stat Soc B 45:311-354, 1983
11. Harrell FE Jr, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic prediction. Stat Med 3:143-152, 1984
12. Spiegelhalter DJ: Probabilistic prediction in patient management and clinical trials. Stat Med 5:421-433, 1986
13. Miller ME, Hui SL: Validation techniques for logistic regression models. Stat Med 10:1213-1226, 1991
14. Chatfield C: Model uncertainty, data mining and statistical inference. J R Statist Soc 158:419-466, 1995
15. Pitkänen O, Niskanen M, Rehnberg S, et al: Intra-institutional prediction of outcome after cardiac surgery: comparison between a locally derived model and the EuroSCORE. Eur J Cardiothorac Surg 18:703-710, 2000
16. Steyerberg EW, Vergouwe Y, Keizer HJ, et al: Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. Stat Med 20:3847-59, 2001
17. van Houwelingen JC, le Cessie S: Predictive value of statistical models. Stat Med 9:1303-1325, 1990
18. Vergouwe Y, Steyerberg EW, de Wit R: External validity of a prediction rule for residual mass histology in testicular cancer: an evaluation for good prognosis patients. Br J Cancer 88:843-847, 2003
19. Cox DR: Two further applications of a model for binary regression. Biometrika 45:562-565, 1958
20. Lemeshow S, Hosmer DW: Applied logistic regression. New York, Wiley, 1989
21. Arkes HR, Dawson NV, Speroff T, et al: The covariance decomposition of the probability score and its use in evaluating prognostic estimates. Med Decis Making 15:120-131, 1995
22. Nagelkerke NJD: A note on the general definition of the coefficient of determination. Biometrika 78:691-692, 1991
23. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36, 1982
24. Hosmer DW, Hosmer T, le Cessie S, et al: A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 16:965-980, 1997

25. Peduzzi P, Concato J, Feinstein AR, et al: Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 48:1503-10, 1995
26. Timsit JF, Fosse JP, Troche G, et al: Calibration and discrimination by daily Logistic Organ Dysfunction scoring comparatively with daily Sequential Organ Failure Assessment scoring for predicting hospital mortality in critically ill patients. Crit Care Med 30:2003-13, 2002
27. Roche N, Herer B, Roig C, et al: Prospective testing of two models based on clinical and oximetric variables for prediction of obstructive sleep apnea. Chest 121:747-52, 200227.
28. Oostenbrink R, Moons KG, Donders AR, et al: Prediction of bacterial meningitis in children with meningeal signs: reduction of lumbar punctures. Acta Paediatr 90:611-617, 2001
29. Culine S, Kramar A, Saghatchian M, et al: Development and validation of a prognostic model to predict the length of survival in patients with carcinomas of an unknown primary site. J Clin Oncol 20:4679-83, 2002
30. Wang Y, Lim LL, Levi C, et al: A prognostic index for 30-day mortality after stroke. J Clin Epidemiol 54:766-73, 2001
31. Gibson RM, Stephenson GC: Agressive management of severe closed head injury: time for reappraisal. Lancet ii:369-371, 1989
32. Feldman Z, Contant CF, Robertson CS, et al: Evaluation of the Leeds prognostic score for severe head injury. Lancet 337:1451-3, 1991
33. Centor RM, Yarbrough B, Wood JP: Inability to predict relapse in acute asthma. N Engl J Med 310:577-80, 1984
34. Woo KS, Pun CO, Wang RY, et al: Validation of a coronary prognostic index for the Chinese--a tale of three cities. Int J Cardiol 23:173-8, 1989
35. Oliver D, Britton M, Seed P, et al: Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. BMJ 315:1049-53, 1997