ELSEVIER

Journal of Clinical Epidemiology 58 (2005) 475–483

Journal of Clinical Epidemiology

Substantial effective sample sizes were required for external validation studies of predictive logistic regression models

Yvonne Vergouwe*, Ewout W. Steyerberg, Marinus J.C. Eijkemans, J. Dik F. Habbema

Department of Public Health, Erasmus MC, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands

Accepted 21 June 2004

Abstract

Background and Objectives: The performance of a prediction model is usually worse in external validation data compared to the development data. We aimed to determine at which effective sample sizes (i.e., number of events) relevant differences in model performance can be detected with adequate power.

Methods: We used a logistic regression model to predict the probability that residual masses of patients treated for metastatic testicular cancer contained only benign tissue. We performed standard power calculations and Monte Carlo simulations to estimate the numbers of events that are required to detect several types of model invalidity with 80% power at the 5% significance level.

Results: A validation sample with 111 events was required to detect that a model predicted too high probabilities, when predictions were on average 1.5 times too high on the odds scale. A decrease in discriminative ability of the model, indicated by a decrease in the *c*-statistic from 0.83 to 0.73, required 81 to 106 events, depending on the specific scenario.

Conclusion: We suggest a minimum of 100 events and 100 nonevents for external validation samples. Specific hypotheses may, however, require substantially higher effective sample sizes to obtain adequate power.

Keywords: External validation; Performance; Prediction models; Sample size; Simulations

1. Introduction

Predictive logistic regression models are important tools to provide estimates of patient outcome probabilities. A model that accurately predicts probabilities for patients in the development data may unfortunately not do so for new patients, even when the patients are derived from plausibly related populations, for example, patients treated more recently or patients from another center [1]. Therefore, the performance of prediction models needs to be tested in new patients (external validation) [2,3]. 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 [1,4,5]. Similarly, the place aspect can be studied by splitting the data according to treatment centers [6–8].

Validation studies may typically show a systematic deviation of the predicted probabilities or too extreme predicted probabilities [9,10]. A systematic deviation of the probabilities (overall too high or too low), suggests that an important predictor variable was not included in the model [11,12]. If the 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 [13–15]. 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 (1) calibration measures, which study the agreement between observed outcome frequencies and predicted probabilities, (2) discrimination measures, which study the ability of the model to distinguish between patients with different outcomes, and (3) overall performance measures, which incorporate both aspects of calibration and discrimination [16,17]. 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 [9,12,18]. Particularly, the

* Corresponding author. Tel.: + 31 30 250 3001; fax: + 31 30 250 5485.
E-mail address: Y.Vergouwe@UMCUtrecht.nl (Y. Vergouwe).

0895-4356/05/$ – see front matter
doi: 10.1016/j.jclinepi.2004.06.017

use of too small samples may lead to statistically nonsignificant results, while true differences do exist. For binary outcomes, the power is determined by the number of events (or nonevents, if less frequent than events), that is, the effective sample size. For instance, a sample with 821 patients may seem adequate, but an outcome frequency of 1.1% implies that the sample contains only nine events [19]. Such a data set provides little power to test differences in model performance. 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 [6,20]. The model was validated in samples that differed in some way from the development data. Evaluations of power were performed with standard formulas for power calculations and with Monte Carlo simulations. We will show that relatively many events are required to obtain a reasonable power in external validation studies.

2. Data and methods

2.1. Prediction model for metastatic testicular germ cell cancer

Resected retroperitoneal lymph nodes of patients treated with chemotherapy for metastatic testicular germ cell cancer contain purely 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 [6]. The model contained six predictor variables: three dichotomous variables (normal prechemotherapy levels of the serum tumor markers alpha-fetoprotein [AFP] and human chorionic gonadotropin [HCG], and absence of teratoma elements in the primary tumor [ter]) and three continuous variables with transformation (natural logarithm of the standardized level of prechemotherapy lactate dehydrogenase [ln(LDH)], square root of the maximum mass size after chemotherapy [sqpost], and change in mass size during chemotherapy per 10% [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). The formula for lp_i in the development sample was: $lp_i = -0.98 + 0.87 * \text{AFP}_i + 0.76 * \text{HCG}_i + 0.86 * \text{ter}_i + 0.97 * \ln(\text{LDH})_i - 0.28 * \text{sqpost}_i + 0.15 * \text{change10}_i$, with dichotomous variables coded 0/1 and continuous variables indicated as above.

2.2. Simulation of validation samples

2.2.1. Monte Carlo simulation

The prediction model for metastatic testicular germ cell cancer was developed with data of 544 patients, of whom 245 had benign tissue (outcome frequency 45%). We used the predictor values (i.e., covariate patterns) of these patients to simulate the validation samples.

To generate Monte Carlo simulated validation samples, covariate patterns were randomly drawn from the development data set with replacement. For each covariate pattern X_i, lp_i was calculated with the prediction model for metastatic testicular germ cell cancer. The outcome value (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 $\pi(X_i) \geqslant u_i$ and $Y_i = 0$ otherwise. This resulted in validation samples similar to the development data set, which mimics the situation that the validation sample originates from the same underlying population as the development data set.

To simulate validation 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 prediction model was studied by comparing $Y.new_i$ with lp_i. This mimics the situation that the original model is tested in a validation sample, 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. Sample sizes were 100, 200, or 500 with 45, 90, and 225 events, respectively, at an outcome frequency of 45%.

2.2.2. Simulated scenarios

The following four scenarios were simulated (Table 1). First, the omission of an important predictor was studied. Predicted probabilities were simulated to be systematically too high by subtracting a constant factor a from the linear predictor; $lp.new_i = lp_i - a$, with $a = \ln(1.5)$ (Scenario I a) or $a = \ln(2)$ (Scenario I b). This corresponds with predictions of the validation sample being 1.5 or two times too high on the odds scale. When predictions are on average too high, the observed frequencies in the validation samples are lower than in the development set. We therefore expect outcome frequencies lower than 45% in the simulated validation samples. Second, the situation of overoptimism was studied, which leads to too extreme predictions. Overoptimistic models are the result of insufficient shrinkage of the regression coefficients, which occurs especially in relatively small data sets [2,21,22]. Too extreme predictions were simulated by multiplying the original linear predictor with a shrinkage factor s; $lp.new_i = s * lp_i$, with $s = 0.8$ (Scenario II a) or 0.6 (Scenario II b). The intercept of the linear predictors used for the simulations were adjusted such that the average outcome frequency remained 45%. Third, we studied a situation in which predictor definitions were different or estimates were imprecise leading to wrong regression coefficients for the validation samples. To simulate this scenario (Scenario III), we used the coefficients as estimated for another population (EORTC/MRC trial) [23]. The $lp.new$

Table 1
Simulated clinical scenarios to study the power of performance measures in external validation studies for a prediction model in metastatic testicular germ cell cancer

No.	Scenario	Results in validation samples	Monte Carlo simulation	
			Technique	Case-mix
0	Model was adequate	Predictions reliable	lp_{new} = lp	All patients
I a	An important predictor was omitted from the model	Predictions systematically too high (or too low)	lp_{new} = lp - ln(1.5)	All patients
I b			lp_{new} = lp - ln(2)	All patients
II a	Internal validation and subsequent shrinkage of regression coefficients was insufficient	Predictions too extreme (low predictions too low, high predictions too high)	lp_{new} = 0.8*lp	All patients
II b			lp_{new} = 0.6*lp	All patients
III	Differences in predictor definitions (modelling strategy adequate), or imprecise estimates (e.g., development dataset too small)	Different regression coefficients	lp_{new} = X_i*coefficients as estimated in an EORTC/MRC trial	All patients
IV	Model was adequate, but validation set contained a more homogeneous patient group	Predictions reliable, but other case-mix	lp_{new} = lp	Subsample of patients with at least three unfavorable predictor values (ter, AFP, change10) for benign tissue

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

was defined as: $lp.new_i = -1.01 + 0.63 * AFP_i + 0.70 * HCG_i + 0.68 * ter_i + 1.04 * \ln(LDH)_i - 0.17 * sqpost_i + 0.03 * change10_i$.

Fourth, a more homogeneous patient group (change in case-mix) was studied using only a subsample of the development data (Scenario IV). We selected patients with unfavorable values for the predictor variables ter (value 0), AFP (value 0), and change10 (value <50%). This simulates a narrowed distribution of probabilities, which affects the discriminative ability of the model. In this scenario, the outcome values were simulated with $lp.new_i = lp_i$, which implies that the calibration was unaffected. The average outcome frequencies in the Scenarios I a, I b, II a, II b, III, and IV were 38, 34, 45, 45, 40, and 10%, respectively.

2.3. Performance measures

The model performance was quantified with respect to calibration and discrimination. Further, the overall performance of the models was quantified.

Calibration, or reliability, refers to the agreement between observed outcome frequencies and predicted probabilities. Calibration was studied with calibration curves, that is, graphic presentations of the relationships between the observed outcome frequencies and the predicted probabilities. Calibration curves can be characterized by a regression line (or calibration line) with intercept (α) and slope (β) [13,17]. These parameters can be estimated in a logistic regression model with the observed outcome as outcome variable and the linear predictor as only predictor variable. 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) [2,13,17]. The Hosmer-Lemeshow statistic for external validity was also estimated. To compute this statistic, predicted probabilities were divided in deciles. Per decile the sum of the observed outcomes were compared with the sum of the predicted probabilities. The statistic has a χ^2-distribution with 10 degrees of freedom in external validation studies [24].

Discrimination refers to the ability to distinguish patients with different outcomes, that is, whether the relative ranking of individual predictions is in the correct order. Discrimination was quantified with the concordance (*c*)-statistic. This statistic is identical to the area under the receiver operating characteristic curve, if the outcome variable is binary [15]. The statistic indicates the proportion of all pairs of patients with different outcome values for which the patient having the outcome has a higher predicted probability than the other patient.

Overall performance measures incorporate both calibration and discrimination aspects. Measures used to quantify the overall performance were the Brier score [25], that is, $\Sigma(Y_i - \pi(X_i))^2/n$ and Nagelkerke's R^2, that is, a measure of explained variation calculated on the log-likelihood scale [26].

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

2.4. Assessment of power

Calibration was studied as the agreement between observed outcomes and predicted probabilities within the same

sample, rather than as a comparison between samples. Therefore, we used one-sample tests for the calibration measures (intercept, slope, *U*-statistic, and Hosmer-Lemeshow statistic). In contrast, estimates of the discriminative ability can be compared between samples. We therefore used two-sample tests to compare the discriminative ability between the development and validation settings.

2.4.1. Standard sample size calculations

The power to detect a statistically significant difference in model performance at a particular sample size was calculated with standard formulas based on the Normal distribution. The number of events is implicitly incorporated in these formulas via the "virtual" standard deviation of the difference in the performance measure (σ). This σ is related to the proportion of events (the outcome frequency). Samples with lower outcome frequencies correspond to a larger σ. For one-sample tests σ equals standard error(performance measure)* $\sqrt{n}$. For two-sample tests σ equals $\sqrt{\sigma_1^2+\sigma_2^2}$, with σ_1 and σ_2 the standard deviations of the performance measure as estimated for the development and validation sample, respectively. When the expected outcome frequency of the validation sample equaled the frequency of the development data set, σ for one sample tests could readily be estimated from the original development data set. Further, σ_2 then equaled σ_1 for two-sample tests. In scenarios with a different expected outcome frequency for the validation samples (I, III, and IV), we used Monte Carlo simulations to estimate σ.

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

$$Z_\beta = \sqrt{\frac{N\delta^2}{\sigma^2}} - Z_{1/2\alpha}$$

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 $1-\beta$ = the power of the hypothesis test, $Z_{1/2\alpha}$ = value of the standard Normal distribution corresponding to $^1/_2\alpha$, with α = type I error rate (here: 0.05).

As an example, we calculate the power of a test for an incorrect intercept of the calibration line. If the model predicts on average 1.5 times too high probabilities on the odds scale ($\delta = \ln(1.5) = 0.405$), we found that the standard error (SE) of the intercept is 0.106 in simulated validation samples of $n = 544$. Then σ = se(intercept)*$\sqrt{n} = 0.106*\sqrt{544} = 2.47$. The power to detect this difference in a validation sample of 100 patients is:

$$Z_\beta = \sqrt{\frac{100[0.405]^2}{2.47^2}} - 1.96 = -0.32$$

corresponding to $\beta = 0.62$ and $1-\beta = 0.38$, or 38% power.

We calculated the power to detect the differences in performance for three sample sizes ($n = 100$, $n = 200$, $n = 500$), and the required sample size to achieve 80% power.

2.4.2. Monte Carlo simulations

We further studied the power nonparametrically with Monte Carlo simulations. These simulations do not make the assumption of a Normal distribution as in the standard power calculation formulas. We considered the same three sample sizes as with the standard power calculations ($n = 100$, $n = 200$, $n = 500$). We could not, however, easily repeat the estimation of the required sample sizes for 80% power. In the simulations, we took the variability of the development data into account for the *c*-statistic by drawing 5,000 development samples of size 544 with replacement. For each sample the optimism-corrected *c*-statistic [2], and its standard error [27] 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 *t*-test for detecting a change in the *c*-statistic was determined by the proportion of two-sided *p*-values below 5% in 5,000 simulated validation samples.

3. Results

3.1. Model performance in validation samples

Fig. 1 shows the calibration curves and performance measures of the prediction model for metastatic testicular germ cell cancer in the simulated validation samples, 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 good discrimination (*c*-statistic = 0.83) and good overall performance ($R^2 = 0.41$, and Brier score = 0.17). Systematically too high predictions in the validation samples (Scenarios I a and I b) showed a change in the intercept of the calibration line, but unchanged discrimination (*c*-statistic remained 0.83). Discrimination was not influenced, because the spread in the observed frequencies remained the same. The overall performance was poorer (R^2 decreased; Brier score increased slightly from 0.167 to 0.168 and 0.172), because of the miscalibration. Too extreme predictions (Scenarios II a and II b) had calibration slopes smaller than 1.0 (0.8 and 0.6). A decrease in discrimination was found (*c*-statistic = 0.79 and 0.74), in line with a decrease in the spread in the observed frequencies. The overall performance decreased more markedly compared with the systematic changes in the predictions of Scenarios I, because both calibration and discrimination were influenced in Scenario II. The scenario with different regression coefficients in the validation samples (Scenario III), showed a reduction in all aspects of model performance. The more homogeneous patient group (Scenario IV) showed in particular a decrease in discrimination (*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), implying better overall performance. However, the Brier score is lower at a lower outcome frequency for an otherwise similar model.

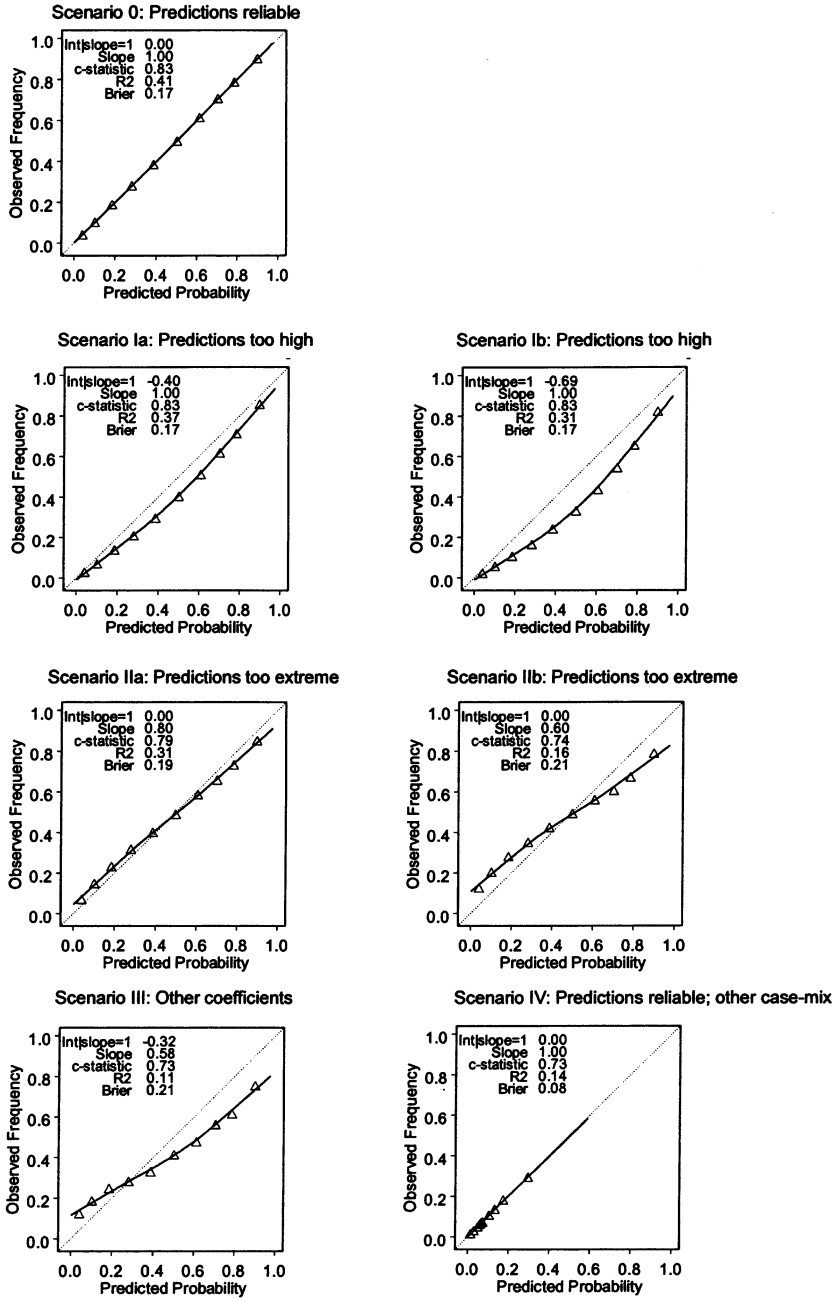

Fig. 1. Calibration curves corresponding to different simulated scenarios (n = 100,000): development data (scenario 0); systematically too high predicted probabilities (I); too extreme probabilities (II); different regression coefficients (III); different case-mix (IV). The dotted line indicates perfect calibration, that is, 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.

Because the samples of Scenario IV contained less patients with benign tissue (outcome frequency 10% vs. 45% in Scenario 0), the positive influence of the low outcome frequency on the Brier score was larger than the negative influence of the reduction in discrimination.

3.2. Power

3.2.1. Standard sample size calculations

Table 2 shows the power of some calibration measures and the t-test for detecting a change in the c-statistic as calculated with standard formulas at three different sample sizes (n = 100, 200, and 500). Also, the required sample sizes to achieve 80% power are shown. We illustrate the power for performance measures that clearly changed in the different scenarios, as shown in Fig. 1. Predictions that were on average 1.5 times too high on the odds scale (Scenario I a) could be detected with the intercept of the calibration line in 38, 64, and 96% of samples of size 100, 200, and 500, respectively. A sample size of n = 292 was required to achieve 80% power (111 events). Too extreme predictions leading

Table 2
Power and sample size calculations to detect a decrease in model performance for a prediction model in metastatic testicular germ cell cancer

Measure	Scenario	Original value	SE_1[a]	New value	SE_2[b]	Power			Required sample size for 80% power (events)	
						$n = 100$	$n = 200$	$n = 500$		
Intercept|Slope=1[c]	I a	0		-ln(1.5)	0.106	38	64	96	292	(111)
	I b	0		-ln(2.0)	0.108	79	97	100	104	(35)
Slope	II a	1		0.8	0.081	19	33	66	694	(312)
	II b	1		0.6	0.068	71	95	100	125	(56)
c-statistic[d]	III	0.83	0.017	0.73	0.022	46	71	94	265	(106)
	IV	0.83	0.017	0.73	0.088	20	33	64	807	(81)

[a] Standard error at outcome frequency of development data.
[b] Standard error at outcome frequency of validation sample.
[c] Value of the intercept, when the value of the slope was set to 1.
[d] Two-sample test.

to a calibration slope of 0.8 (Scenario II a) could be detected in only 66% of all samples of size 500 (225 events). To achieve 80% power, a sample of size 694 with 312 events would be required. A calibration slope of 0.6 (Scenario II b) could be detected with 80% power in samples of size 125 (56 events). A decrease in the *c*-statistic from 0.83 to 0.73 at an outcome frequency of 40% (Scenario III) could be well detected in 71% of the samples of 200 patients with 80 events. At an outcome frequency of 10% (Scenario IV), a sample of 807 patients with 81 events would be required to achieve 80% power.

3.2.2. Monte Carlo simulations

Fig. 2 shows the power of all calibration measures considered and the *t*-test for detecting a change in the *c*-statistic at three different sample sizes ($n = 100$, $n = 200$, and $n = 500$) as estimated with the simulated validation samples. The results confirmed the results of the standard formulas in Table 2.

The Null hypothesis of equal model performance in the validation samples was rejected in around 5% of the samples of Scenario 0, in agreement with the nominal significance level of 5%. The power to detect an intercept problem in a model with on average 1.5 times too high predictions (Scenario I a) was in agreement with Table 2 (37; 64; and 96% vs. 38, 64, and 96%). The Hosmer-Lemeshow statistic had much less power to detect an intercept problem, that is, 26% in samples of 200 subjects compared with 54% power for the *U*-statistic.

Overoptimism as reflected by a slope of 0.8 (Scenario II a) was detected with the calibration slope in 65% of all simulated validation samples of 500 patients with 225 events. Validation samples of 100 patients with 45 events had poor power to detect such overoptimism, with power $\leqslant 20\%$ for the different performance measures in Scenario II a. The large overoptimism simulated in Scenario II b (slope = 0.6) was well detected with the calibration slope, the *U*-statistic, and the Hosmer-Lemeshow statistic in samples of 200 patients (90 events) or more. The actual type I error was 6.3–9.5% for a change in intercept.

All measures had reasonable or good power to detect lack of performance if the regression coefficients were different as in Scenario III. The power for the *t*-test to detect the decrease in *c*-statistic (from 0.83 to 0.73) in Scenario IV was lower (13, 26, and 68%) than the power in Scenario III (48, 77, and 97%), while the decrease in *c*-statistic was similar (around 0.1). This was mainly caused by the lower average outcome frequency (10 vs. 40%), resulting in different number of events (10, 20, and 50 vs. 40, 80, and 200, respectively). The power became comparable when we increased the average outcome frequency to 40% in Scenario IV.

4. Discussion

We have shown that substantial sample sizes and numbers of events are required to detect relevant decreases in model performance in external validation samples. We studied several measures for calibration and discrimination. A model showing systematically too high predictions in the validation sample could be best detected with a test for the intercept of the calibration line. This measure showed far more power than the Hosmer-Lemeshow goodness-of-fit statistic. The *U*-statistic tests the values of intercept and calibration slope jointly [13,17] and had intermediate power. Too extreme predictions (overoptimism) were best detected with the calibration slope, followed by the *U*-statistic, the Hosmer-Lemeshow statistic, and the *t*-test to detect a change in the *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 validation samples with higher number of events, reflected by higher average outcome frequencies.

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 [24]. It has been shown that this statistic had poor power for various violations of model assumptions, and reasonable power to detect whether nonlinear terms or alternative link

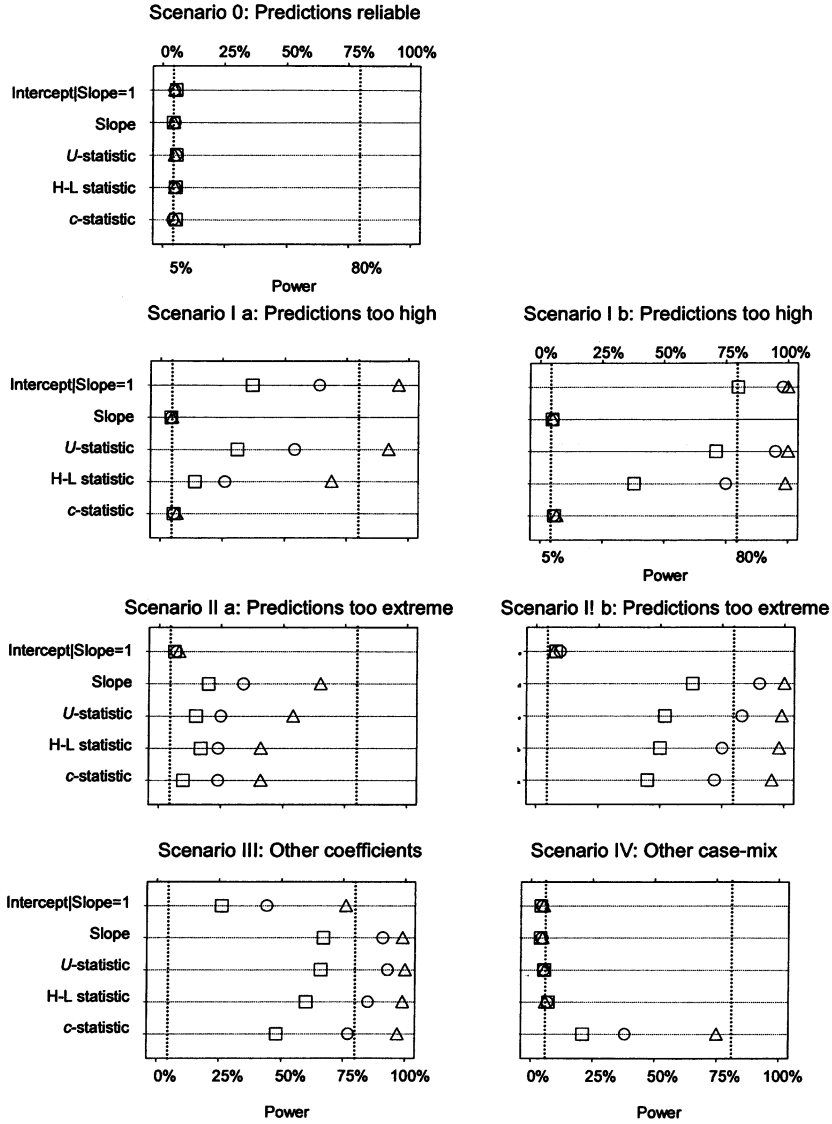

Fig. 2. Power in percentages to detect differences in performance of a prediction model for several clinical scenarios at three sample sizes (square, $n = 100$; circle, $n = 200$; and triangle, $n = 500$). The nominal Type I error was set at 0.05 (first dotted line). Power was considered adequate at 80% (second dotted line) or higher.

functions are required [28]. Therefore, the Hosmer-Lemeshow statistic may better be reserved for model development and not used for validation purposes. A good alternative in model validation may be the *U*-statistic, which tests the intercept and calibration slope jointly [2,13,17].

The power calculations and the Monte Carlo simulations showed similar results. This indicates that the assumed Normal distributions of the performance measures were reasonable. The power can hence easily be calculated, under the assumption that the standard deviation ("virtual" σ) estimated in the development data equals the σ of the validation sample. However, σ depends on the average outcome frequency, with larger σ when the expected outcome frequency in the validation sample is lower than the outcome frequency in the development data. In such situations, simulation studies remain valuable. Furthermore, 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 of these measures, 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 to evaluate systematically within each of the simulated validation samples with current 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.

For instance, Scenario IV showed the influence of the average outcome frequency on the power to detect differences in performance. A decrease in the *c*-statistic from 0.83 to 0.73 was detected at an average outcome frequency of 10% (Scenario IV) in only 38% of the validation samples containing 200 patients. The power to detect the decrease in the *c*-statistic was 88% at an average outcome frequency of 45% in validation samples of size 200 and the same homogeneity as in Scenario IV (data not shown). Thus, studying a decrease in model performance for populations with much lower or higher average outcome frequencies will require different sample sizes. These findings confirm that the effective sample size is determined by the number of events (the average outcome frequency times the sample size) or nonevents (in case of a frequent outcome), rather than the total number of subjects [12,15,21,29].

The number of events (or nonevents) also sets limits to model development. Harrell and others proposed to use only one degree of freedom per 10 events for model development [15,22,29]. A more unfavorable ratio might result in a poorly validating model [2]. In accordance with the 1:10 rule, we can formulate a rule of thumb for the minimum number of events and nonevents required to properly study model performance in new data. We found that samples with around 100 events had approximately 80% power to detect substantial differences in model performance such as 1.5 times too high or low predicted odds (111 events), or a decrease in the *c*-statistic of 0.1 (81 to 106 events). A rule of thumb may hence be to use a minimum of 100 events and 100 nonevents when studying the model performance in a new population. Smaller samples only have power to detect quite large differences in model performance, such as a decrease in the calibration slope to 0.6 (56 events were required). On the other hand, larger validation samples would be required to detect smaller differences in performance.

In the literature, sample sizes of validation sets differ over a wide range [30,31]. 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 [32–34]. A review [9] described validation samples with sizes varying between 52 and 479 patients. The numbers of events ranged from 24 to 115 [35–39], implying that most of these studies were underpowered to detect relevant differences ($<$100 events).

There are several limitations to our study. First, the prediction model for metastatic testicular germ cell cancer is only one example, and results may vary for other prediction models. Also, we started with simulations that assume that all logistic model assumptions hold, which will never be fully the case in empirical research [18]. Further, the chosen clinical scenarios for which the power was studied, are arbitrary. We consider the results informative because 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. 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. Such more subtle model violations may well require larger validation sample sizes to be detected.

Furthermore, the validity of a prediction model in a new population also depends 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, is higher for a large development sample [12]. However, we did not take this development uncertainty into account for the calibration measures, which were tested with one-sample tests. Our reasoning was that a literature model that is to be adopted by others will often be considered as fixed. The uncertainty in the predicted probabilities is ignored then, leading to smaller sample sizes than when this uncertainty was incorporated. In contrast, the uncertainty in estimates from the development sample was taken into account in studying discrimination, because we may want to compare discrimination between the development and validation setting.

In conclusion, we recommend to estimate both the intercept and slope of the calibration line when externally validating a model. Tests for these measures showed the highest power to detect miscalibration. The *U*-statistic tests the intercept and slope jointly and can be used as an overall calibration test. A recent simulation study suggests that such recalibration parameters may well be used to provide updated model predictions [12]. We note that aiming for updating a prediction model may even be more sensible than testing its validity. We further suggest to aim for validation samples with at least 100 events and 100 nonevents. Validation studies that did not show clear differences in performance might actually have been underpowered, because often much less than 100 events were included. Valuable information may in such situations be obtained from confidence intervals of performance measures, which would be quite wide in small validation studies.

Acknowledgments

This manuscript was substantially improved by the thoughtful comments of a reviewer. Yvonne Vergouwe was supported by The Netherlands Organization for Scientific Research. Ewout Steyerberg was supported by a fellowship from the Royal Netherlands Academy of Arts and Sciences.

References

[1] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.

[2] 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 1996;15:361–87.
[3] Hand JD. Construction and assessment of classification rules. Chisester, England: John Wiley & Sons Ltd.; 1997.
[4] Picard RR, Berk KN. Data splitting. Am Stat 1990;44:140–7.
[5] Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Making 1993;13: 49–58.
[6] Steyerberg EW, Keizer HJ, Fosså SD, Sleijfer DT, Toner GC, Schraffordt Koops H, Mulders PF, Messemer JE, Ney K, Donohue JP, Bajorin DF, Stoter G, Bosl GJ, Habbema JD. Prediction of residual retroperitoneal mass histology following chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. J Clin Oncol 1995;13:1177–87.
[7] van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med 1995;14:1999–2008.
[8] Krijnen P, van Jaarsveld BC, Steyerberg EW, Man in't Veld AJ, Schalekamp MA, Habbema JD. A clinical prediction rule for renal artery stenosis. Ann Intern Med 1998;129:705–11.
[9] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73.
[10] Terrin N, Schmid CH, Griffith JL, D'Agostino RB, Selker HP. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. J Clin Epidemiol 2003; 56:721–9.
[11] Vergouwe Y, Steyerberg EW, Foster RS, Habbema JD, Donohue JP. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. J Urol 2001;165:84–8.
[12] Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004; 23:2567–86.
[13] Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45:562–5.
[14] Copas JB. Regression, prediction and shrinkage. J R Stat Soc B 1983;45:311–54.
[15] Harrel FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143–52.
[16] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.
[17] Miller ME, Hui SL. Validation techniques for logistic regression models. Stat Med 1991;10:1213–26.
[18] Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc A 1995;158:419–66.
[19] Pitkänen O, Niskanen M, Rehnberg S, Hippelainen M, Hynynen M. Intra-institutional prediction of outcome after cardiac surgery: comparison between a locally derived model and the EuroSCORE. Eur J Cardiothorac Surg 2000;18:703–10.
[20] Steyerberg EW, Vergouwe Y, Keizer HJ, Habbema JD. Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. Stat Med 2001;20:3847–59.
[21] van Houwelingen JC, le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303–25.
[22] Steyerberg EW, Eijkemans MJC, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000; 19:1059–79.
[23] Vergouwe Y, Steyerberg EW, de Wit R, Roberts JT, Keizer HJ, Collette L, Stenning SP, Habbema JDF. External validity of a prediction rule for residual mass histology in testicular cancer: an evaluation for good prognosis patients. Br J Cancer 2003;88:843–7.
[24] Lemeshow S, Hosmer DW. Applied logistic regression. New York: Wiley; 1989.
[25] Arkes HR, Dawson NV, Speroff T, Harrell FE Jr, Alzola C, Philips R, Desbiens N, Oye RK, Knaus W, Connors AF Jr. The covariance decomposition of the probability score and its use in evaluating prognostic estimates. Med Decis Making 1995;15:120–31.
[26] Nagelkerke NJ. A note on the general definition of the coefficient of determination. Biometrika 1991;78:691–2.
[27] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143: 29–36.
[28] Hosmer DW, Hosmer T, le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16:965–80.
[29] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503–10.
[30] Timsit JF, Fosse JP, Troche G, De Lassence A, Alberti C, Garrouste-Orgeas M, Bornstain C, Adrie C, Cheval C, Chevret S. 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 2002;30:2003–13.
[31] Roche N, Herer B, Roig C, Huchon G. Prospective testing of two models based on clinical and oximetric variables for prediction of obstructive sleep apnea. Chest 2002;121:747–52.
[32] Oostenbrink R, Moons KG, Donders AR, Grobbee DE, Moll HA. Prediction of bacterial meningitis in children with meningeal signs: reduction of lumbar punctures. Acta Paediatr 2001;90:611–7.
[33] Culine S, Kramar A, Saghatchian M, Bugat R, Lesimple T, Lortholary A, Merrouche Y, Laplanche A, Fizazi K. 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 2002;20: 4679–83.
[34] Wang Y, Lim LL, Levi C, Heller RF, Fischer J. A prognostic index for 30-day mortality after stroke. J Clin Epidemiol 2001;54:766–73.
[35] Gibson RM, Stephenson GC. Agressive management of severe closed head injury: time for reappraisal. Lancet 1989;334:369–71.
[36] Feldman Z, Contant CF, Robertson CS, Narayan RK, Grossman RG. Evaluation of the Leeds prognostic score for severe head injury. Lancet 1991;337:1451–3.
[37] Centor RM, Yarbrough B, Wood JP. Inability to predict relapse in acute asthma. N Engl J Med 1984;310:577–80.
[38] Woo KS, Pun CO, Wang RY, Ma H, Huang ZZ, Dai RH, Huang DJ, Vallance-Owen J. Validation of a coronary prognostic index for the Chinese—a tale of three cities. Int J Cardiol 1989;23:173–8.
[39] Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Developmentment and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case–control and cohort studies. BMJ 1997;315:1049–53.