ELSEVIER

Journal of Clinical Epidemiology 56 (2003) 441–447

Journal of Clinical Epidemiology

Internal and external validation of predictive models: A simulation study of bias and precision in small samples

Ewout W. Steyerberg[a,*], Sacha E. Bleeker[b], Henriëtte A. Moll[b], Diederick E. Grobbee[c], Karel G.M. Moons[c]

[a]*Center for Clinical Decision Sciences, Department of Public Health, Erasmus MC, PO Box 1738 3000 DR, Rotterdam, The Netherlands*
[b]*Outpatient Department of Pediatrics, Sophia Children's Hospital, Dr Molewaterplein 60, 3015 GJ Rotterdam, The Netherlands*
[c]*Julius Center for Health Sciences and Primary Care, University Medical Center, PO Box 80035, 3508 TA, Utrecht, The Netherlands*

Received 28 May 2002; received in revised form 14 December 2002; accepted 10 January 2003

Abstract

We performed a simulation study to investigate the accuracy of bootstrap estimates of optimism (internal validation) and the precision of performance estimates in independent validation samples (external validation). We combined two data sets containing children presenting with fever without source ($n = 376 + 179 = 555$; 120 bacterial infections). Random samples were drawn from this combined data set for the development ($n = 376$) and validation ($n = 179$) of logistic regression models. The models included statistically significant predictors for infection selected from a set of 57 candidate predictors. Model development, including the selection of predictors, and validation were repeated in a bootstrapping procedure. The resulting expected optimism estimate in the receiver operating characteristic (ROC) area was compared with the observed optimism according to independent validation samples. The average apparent ROC area was 0.74, which was expected (based on bootstrapping) to decrease by 0.07 to 0.67, whereas the observed decrease in the validation samples was 0.09 to 0.65. Omitting the selection of predictors from the bootstrap procedure led to a severe underestimation of the optimism (decrease 0.006). The standard error of the observed ROC area in the independent validation samples was large (0.05). We recommend bootstrapping for internal validation because it gives reasonably valid estimates of the expected optimism in predictive performance provided that any selection of predictors is taken into account. For external validation, substantial sample sizes should be used for sufficient power to detect clinically important changes in performance as compared with the internally validated estimate.

Keywords: Prediction models; Internal validation; Bootstrap; External validation; Logistic Regression

1. Introduction

Optimism is a well-known problem of predictive models: Their performance in new patients is often worse than expected based on performance estimated from the development data set ("apparent performance") [1–3]. The extent of optimism of pre-specified models can be estimated for similar patient populations using internal validation techniques such as bootstrapping [4,5]. Predictive models are, however, usually not pre-specified but are constructed in an iterative way. Model specification may include decisions on coding of variables (e.g., [re-]grouping categorized or continuous variables) and decisions on the inclusion of main effect, nonlinear, and interaction terms in the final model. However, when model specification, such as stepwise selection of predictor variables, can be formulated in a systematic way, it may be replayed entirely in every bootstrap sample [6]. Such a procedure should provide an honest estimate of the optimism of the final model [3,7]. Because empirical evidence for this claim is limited, our first aim was to study the accuracy of the bootstrap estimate of optimism of a prediction model that is developed using variable selection techniques.

Of more interest than internal validity is external validity, or generalizability [8]. External validity is typically studied in independent validation samples with patients from a different but "plausibly related" population [9]. In a previous study, a diagnostic model developed to estimate the presence of a serious bacterial infection in children with fever without apparent source showed a surprisingly poor external validity in another sample of 179 children [10]. Although a sample size of 179 subjects is not uncommon in diagnostic (validation) research, it raises the question of how large a validation set needs to be. Our second aim was to study the precision of

* Corresponding author. Tel.: +31 10 408 7053; fax: +31 10 408 9455.
E-mail address: E.Steyerberg@ErasmusMC.nl (E.W. Steyerberg).

0895-4356/03/$ – see front matter
doi: 10.1016/S0895-4356(03)00047-7

performance estimates in relatively small validation samples and to explore the consequences for the power of validation studies when comparing model performance between development and validation sets.

2. Methods

2.1. Patients

We combined two previously described data sets of children presenting with fever without apparent source: a development set from Rotterdam, The Netherlands, diagnosed between 1988 and 1992 ($n = 376$), and a validation set from Rotterdam and The Hague, diagnosed between 1997 and 1998 ($n = 179$) [10]. Of these 555 children, 120 (22%) had a serious bacterial infection, which was defined as the presence of bacterial meningitis, sepsis or bacteriemia, pneumonia, urinary tract infection, bacterial gastroenteritis, osteomyelitis, or ethmoiditis. In the present analyses, we discarded the original distinction according to time and place [10]. Instead, we drew random samples of $n = 376$ for model development, leaving $n = 179$ for model validation. The average prevalence of infection was 22% in the development and validation samples but could vary due to chance. The random sampling leads to the generation of two study samples under the null hypothesis that no systematic differences exist. This is in contrast to the real-life situation, where true differences may exist between development and validation samples. The sample sizes of 376 and 179 are similar to cross-validation, with two thirds of a data set for model development and one third for model validation [11].

2.2. Model development strategy

In each development sample, logistic regression models were fitted. The presence of a serious bacterial infection was the binary outcome. The models included statistically significant predictors ($P < .05$) selected from a set of 57 candidate predictors. Selection was based on univariable selection followed by a forward stepwise multivariable selection. Selection started with the predictor with the lowest univariable P value, and predictors were added until none of the remaining univariably selected predictors (with $P < .05$) had a multivariable $P < .05$ according to a likelihood ratio test. In our previous study, we used slightly more liberal selection criteria: $P < .15$ for univariable selection and $P < .10$ for multivariable selection [10].

2.3. Model performance

We studied the receiver operating characteristic (ROC) area as the primary indicator of the model performance. The ROC curve was constructed by calculating the sensitivity and specificity for consecutive cut-off points according to the predicted probabilities from the logistic regression models. In addition, we studied the amount of variance explained by the model (R^2) on the log-likelihood scale [12]. Although R^2 is not often used for logistic regression models, it is an attractive measure because it ranges from 0% to 100%. Furthermore, it evaluates model performance on the log likelihood scale, which is also used to fit the model. Statistically, the ROC area is related to Somer's D, which is an example of a rank order statistic, whereas R^2 is an example of a logarithmic scoring rule for model-based predictions [3].

We evaluated the calibration of model predictions by the slope of the linear predictor of the logistic regression model [13]. The linear predictor was defined as the sum on the log scale of the regression coefficients (estimated in the development sample) multiplied by the patient value of the corresponding predictor. This calibration slope is, by definition, unity in the development set. In a validation set, the calibration slope can be estimated using a logistic regression model with the linear predictor as the sole covariate. The slope is generally smaller than 1, reflecting a need for shrinkage of the regression coefficients that were estimated in the development sample [2,3,14,15]. In the context of optimism, we therefore refer to the calibration slope as the "shrinkage factor." The shrinkage factor may be used for multiplication with the originally estimated regression coefficients such that better calibrated predictions are obtained for future patients [2].

2.4. Bias in expected optimism

We studied the expected optimism of the three performance measures (ROC area, R^2, and calibration slope) that were estimated in the development sample with a bootstrap resampling procedure. The bias of the expected optimism was determined by comparison with the observed optimism as was obtained from the validation samples.

For the expected optimism, a random bootstrap sample was drawn with replacement from each development sample ($n = 376$) [4]. The model development strategy was fully followed within the bootstrap sample, including the univariable and multivariable selection. The expected optimism was calculated from the differences between the performance of models developed in each of the bootstrap samples and their performance in the development sample [3]. The bootstrap sample and the development sample are not independent; rather, they contain some of the same patients. On average, 63.2% of the patients from the development sample are included at least once in a bootstrap sample [4]. Variants of the bootstrap procedure have been proposed that consider only independent patients for model validation [16], but in an earlier study these variants had no advantage over the simpler procedure described above [5].

The observed optimism in the model performance measures was defined as the difference between the apparent performance (i.e., before bootstrapping) in the development sample ($n = 376$) and the performance in the independent validation sample ($n = 179$). The expected and

observed optimism should be identical if the bootstrap procedure gives unbiased estimates.

We performed 1000 simulations with a calculation of the expected optimism (model development in one bootstrap sample with model testing on the development sample) and observed optimism (model development in the development sample with validation in an independent sample). Simulations were performed using S-plus software (version 2000; Mathsoft, Cambridge, MA). The mean performance and optimism were calculated with 10% trimming to improve stability.

2.5. Precision of model performance estimates

Bootstrap estimates of expected optimism in model performance become more precise with higher numbers of bootstrap repetitions. At least 100 bootstraps are usually recommended [4,17]. In our previous study, we used 200 bootstrap repetitions, assuming a stable estimate with this number [10]. In the present analyses, we first used the full data set ($n = 555$) to illustrate the variability of the bootstrap resampling procedure with increasing numbers of repetitions.

The precision of the expected optimism and optimism-corrected performance estimates depends not only on the bootstrap variability but also on the sampling variability of the development sample. The bootstrap variability can go to zero with infinite bootstrap repetitions, but the sampling variability remains for a given finite sample size [4]. Therefore, we subsequently evaluated the influence of the sampling variability by randomly drawing 100 development samples from the full data set. We reasoned that 100 development samples was sufficient for estimation of sampling variability. Within each development sample, we performed 200 bootstrap repetitions to make the bootstrap variability of minor importance compared with the sampling variability. In total, 20,000 logistic regressions were performed.

2.6. Power for differences in model performance

The power of a test for a difference in performance was studied for combinations of a development data set with 376 patients and three sizes of the validation set ($n = 179$, $n = 376$ [$\approx 2.1 \times 179$], and $n = 752$ [2×376]), all with a 22% average frequency of the outcome. Hypothetical differences between internally validated and externally validated performance were postulated, assuming that the empirical SEs were known. First, we considered standard two-sample t tests, with two-sided $\alpha = 0.05$ for statistical significance between the internally validated (expected) performance and externally observed performance. Second, we assessed external validation as a one-sample problem (i.e., considering the expected performance fixed). Also, a one-sided test might be appropriate because we might be interested only in the statistical significance of a decrease in model performance in an external validation sample.

3. Results

3.1. Expected optimism in the full data set

In the full data set of 555 children, four statistically significant predictors were selected after univariable and multivariable stepwise analyses: duration of fever at presentation (days), presence of chest-wall retractions, poor peripheral circulation, and presence of crepitations. The apparent ROC area was 0.727, the R^2 was 15.7%, and the calibration slope was unity (Table 1).

According to 1000 bootstrap samples, the expected optimism was 0.056 for the ROC area (0.761 − 0.706) and 9.8% (24.0%−14.2%) for R^2. The shrinkage factor was 0.71. Fig. 1 illustrates the variability of these optimism estimates. The expected optimism estimates were reasonably precise after 200 bootstraps (e.g., for the ROC area the optimism

Table 1
Bootstrap results for the full data set and random development sets[a]

	Full data set ($n = 555$), mean ± SE_B	Random development sets ($n = 376$), mean ± SE_{Sim}
Apparent performance		
ROC area	0.727	0.737 ± 0.035
R^2	15.7%	19.2% ± 4.6%
Calibration slope	1	1
Bootstrap performance		
ROC area	0.761 ± 0.040	0.777 ± 0.017[b]
R^2	24.0% ± 6.2%	27.0% ± 2.7%[b]
Calibration slope	1	1
Test performance[c]		
ROC area	0.706 ± 0.033	0.707 ± 0.019[b]
R^2	14.2% ± 4.2%	13.7% ± 2.6%[b]
Calibration slope	0.71 ± 0.13	0.61 ± 0.057[b]
Expected optimism[d]		
ROC area	0.056 ± 0.025	0.070 ± 0.0095[b]
R^2	9.8% ± 4.7%	13.2% ± 1.3%[b]
Shrinkage factor	0.71 ± 0.13	0.61 ± 0.057[b]
Optimism-corrected performance[e]		
ROC area	0.672	0.665 ± 0.035[b]
R^2	5.9%	5.9% ± 4.7%[b]
Calibration slope	0.71	0.61 ± 0.057[b]

Abbreviations: ROC, receiver operator characteristic; R^2, variance.

[a] The bootstrap procedure leads to estimates of the optimism-corrected performance, which is calculated as apparent performance minus optimism. Means and empirical standard errors (SE, based on standard deviations over samples) are shown. In the full data set, 1000 bootstrap repetitions were used for calculation of both the mean and SE (SE_B). For the random development sets, 1000 simulations were used for the means and for the SE (SE_{Sim}) of the apparent performance. However, for the estimation of the SE of the expected optimism in the random development sets, the evaluation included 200 bootstrap repetitions in only 100 simulations.

[b] SE based on 100 simulations with 200 bootstraps.

[c] The test performance is defined as the performance of the models from the bootstrap samples when applied to the original sample.

[d] The expected optimism was calculated as the difference between bootstrap performance and test performance.

[e] The optimism-corrected performance was defined as apparent performance−optimism.

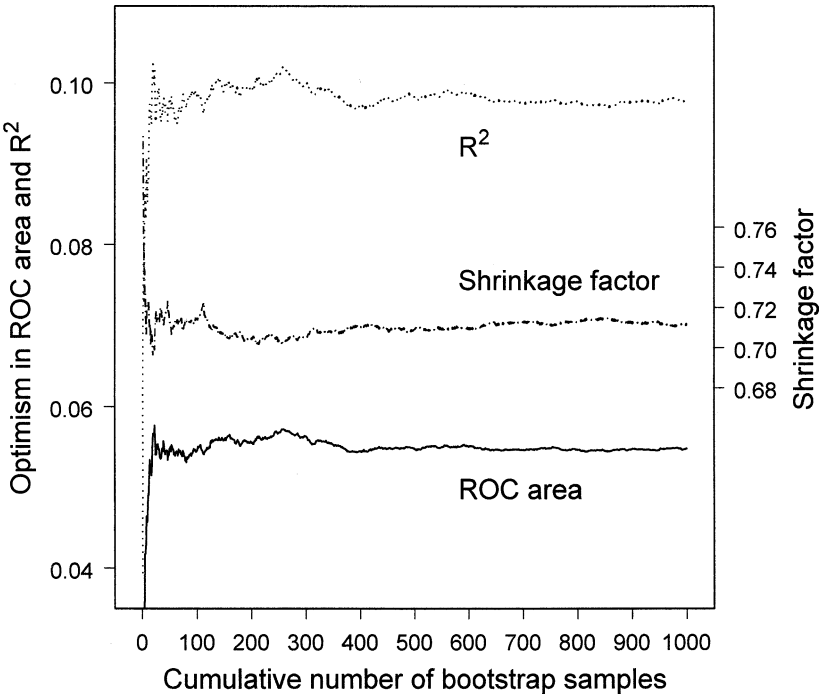

Fig. 1. Cumulative average of the estimated optimism in ROC area and R^2 together with the estimated shrinkage factor in 1000 bootstrap samples consisting of 555 children. The estimates become more stable with increasing number of bootstraps and approach plateau values of 0.056, 9.8%, and 0.71, respectively.

was 0.056 with a 95% confidence interval of 0.052–0.059). The optimism estimates became more stable as more bootstrap samples were added, and a clear plateau was reached after 500 bootstraps. The empirical SEs over the 1000 bootstraps were 0.025, 4.7%, and 0.13 for the ROC area, R^2, and shrinkage factor, respectively (Table 1).

Additionally, we estimated the optimism given the model structure with the four selected predictors in every bootstrap sample as if the model were pre-specified (i.e., ignoring the univariable and multivariable selection process). The optimism was then estimated much smaller as 0.006, 1.9%, and 0.94 for the ROC area, R^2, and shrinkage factor, respectively. Hence, falsely ignoring the selection process led to a substantial underestimation of the optimism.

3.2. Expected optimism in simulated development data sets

In the 1000 randomly drawn development samples with 376 children, on average 4.3 predictors were selected from the set of 57 covariates. The mean of the apparent performance in these 1000 samples was 0.737 for the ROC area and 19.2% for R^2 (Table 1). These apparent performance estimates were higher than for the full sample of 555 children. Based on bootstrapping, the expected optimism was 0.070 (0.777−0.707) for the ROC area and 13.2% (27.0%−13.7%) for R^2. The shrinkage factor was 0.61, which was smaller than for the full sample of 555 children, where the shrinkage factor was 0.71.

3.3. Bias in expected optimism

The expected optimism and optimism-corrected performance as estimated with bootstrapping agreed well with that observed with independent validation samples (Table 2). The expected optimism in ROC area (0.070) was slightly lower than that observed (0.737−0.651 = 0.085), whereas they should ideally be the same. Also, slightly more shrinkage than expected was required (observed calibration slope 0.57 versus expected 0.61). In contrast, the expected optimism in R^2 was slightly too large (expected 13.2% versus observed 11.6%).

3.4. Precision of model performance estimates

The expected optimism could precisely be estimated with 200 bootstraps with minor sampling variability (Table 2). For example, the empirical standard error was only 0.0095 for the expected optimism in ROC area and was 1.3% for R^2. In contrast, the estimates of the optimism-corrected ROC area and corrected R^2 had substantially larger empirical SEs (0.035 and 4.7%, respectively). These SEs were similar to or only slightly larger than the SEs of the apparent performance (0.035 and 4.6%). Hence, the optimism corrections added only minor uncertainty to the estimates of model performance.

The SEs were larger in the validation samples with $n = 179$. For example, the SE of the ROC area was 0.052, and the SE of the calibration slope was 0.23 (Table 2). On closer inspection, we found that the ROC area estimates

Table 2
Accuracy of expected and observed optimism and optimism-corrected performance estimates[a]

	Expected, mean ± SE_{Sim}	Observed, mean ± SE_{Sim}
Optimism[b]		
ROC area	0.070 ± 0.0095[c]	0.085 ± 0.058
R^2	13.2% ± 1.3%[c]	11.6% ± 6.3%
Shrinkage factor	0.61 ± 0.057[c]	0.57 ± 0.23
Optimism-corrected performance[d]		
ROC area	0.665 ± 0.035[c]	0.651 ± 0.052
R^2	5.9% ± 4.7%[c]	7.3% ± 4.6%
Calibration slope	0.61 ± 0.057[c]	0.57 ± 0.23

Abbreviations: ROC, receiver operator characteristic; R^2, variance.

[a] Expected performance was based on random development data sets with $n = 376$, and observed performance was based on random validation data sets with $n = 179$. Means and empirical standard errors (SE, based on standard deviations over samples) are shown. 1000 simulations were used for calculation of the means and SE (SE_{Sim}). However, the calculation of SE of the expected optimism and optimism-corrected performance included 200 bootstrap repetitions in only 100 simulations.

[b] The expected optimism was calculated as the difference between bootstrap performance and test performance (see Table 1). The observed optimism was calculated as the difference between the apparent performance in the random development sets and observed performance in the random validation sets.

[c] SE based on 100 simulations with 200 bootstraps.

[d] The expected optimism-corrected performance was defined as apparent performance−optimism for the random development sets (see Table 1). The observed optimism-corrected performance was equal to the observed performance in the random validation sets.

and calibration slope estimates did reasonably follow a normal distribution (Fig. 2), but the distribution of the estimated R^2 was skewed to the right.

3.5. Power for differences in model performance

Table 3 shows the implications of the substantial SEs in the final row of Table 2 on the power of tests for external validity. With standard two-sided, two-sample t tests, the power would be low, with $n = 376$ for model development and $n = 179$ for model validation. For example, a difference in ROC area by 0.10 (0.75 to 0.65) would be detected as statistically significant in only 36% of the external validation studies. With larger validation sample sizes, the power for such a difference would increase (e.g., to 64% with $n = 752$). The power would also increase when one-sided, one-sample tests were applied (e.g., to 61% for a true difference in ROC area of 0.10) (Table 3).

4. Discussion

We found that internally validated estimates of model performance could accurately be obtained with bootstrapping when a stepwise selection strategy was followed in the construction of the predictive model provided that this strategy was systematically replayed in every bootstrap sample. The expected optimism was close to that observed in independent random validation samples for a number of performance measures, including the ROC area. However, the variability (SEs) of these performance estimates in the validation samples was quite large. This implied a limited power of validation studies to identify clinically important decreases in performance (e.g., a reduction in ROC area by 0.10 on a scale that ranges from 0.5 to 1.0 for sensible models).

The selection strategy in this study consisted of a univariable screening step followed by stepwise multivariable selection among the univariably significant predictors. This strategy is not uncommon. Motivations include the idea that a parsimonious model is preferable to a more complex one and practical arguments. For example, in univariable screening, not all variables have to be considered in the stepwise multivariable analysis. Multivariable stepwise selection led to a simple predictive model (e.g., consisting of only four predictors in our sample of 555 children). However, many disadvantages of stepwise selection methods have been reported, including limited power to select true predictors, the inclusion of noise variables by multiple testing, instability in the selected set of predictors, and bias in the estimated regression coefficients [1,3,7,15,18–20]. These problems hold for forward and backward stepwise selection, although

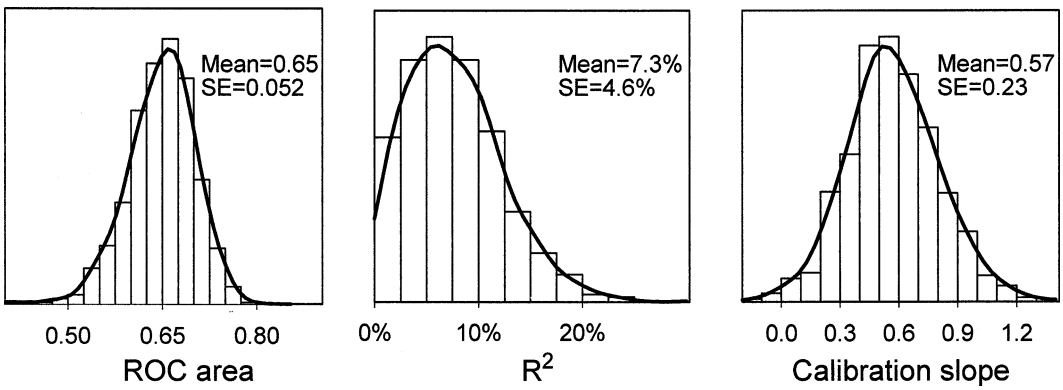

Fig. 2. Distribution of performance estimates in 1000 simulated validation sets with 179 children. We note a substantial variability around the mean performance estimates (see Table 2). Furthermore, the distributions are not Normal (in particular, the distribution of R^2 estimates is skewed to the right).

Table 3
Estimates of power to detect a difference in performance estimates between a development and validation set, given the SEs as found in the simulation study with different hypothetical underlying true differences in performance

Measure	SE in development and validation sets[a]	Hypothetical true performances[a]	Power: 2-sided, 2-sample (n = 376 and n = 179/376/752)[b]	Power: 1-sided, 1-sample (n = 179/376/752)[c]
ROC area	0.035−0.052	0.75−0.70	13%/17%/21%	25%/40%/62%
		0.75−0.65	36%/51%/64%	61%/87%/98.8%
R^2	4.7%−4.6%	15%−10%	12%/14%/16%	29%/47%/71%
		15%−5%	33%/43%/49%	70%/93%/99.7%
Calibration slope	0.057−0.23	0.8−0.6	13%/22%/35%	22%/35%/55%
		0.8−0.5	24%/42%/66%	36%/59%/84%

[a] First number: development set; second number: validation set.
[b] 2-sided, 2-sample t tests were assumed to be performed with $\alpha = 0.05$, given the empirical SE in the development and validation sets.
[c] 1-sided, 1-sample t tests were assmed to be performed with $\alpha = 0.05$, considering only the empirical SE in the validation sets.

backward selection is generally preferable if stepwise selection is applied. The problems are especially evident in relatively small data sets (i.e., with few events), as in our study with 75 events [10]. Because it would be impractical to include all 57 predictors in a predictive model, some form of selection is required [21]. Empirical studies suggest that subject knowledge is of utmost importance (e.g., from formal meta-analysis or qualitative review of previous studies or from clinical experts) [1,15,22,23]. In our case of prediction of serious infections in young children, this approach was not entertained and was not feasible because of the lack of consistent knowledge. However, the chosen stepwise selection approach may have put too much trust in the relatively limited development data set of 376 children with 75 events to identify predictive characteristics among 57 candidate predictors [10].

The expected optimism in model performance was substantial (e.g., a decrease in ROC area from 0.74 to 0.67 and a shrinkage factor of 0.61). Especially dramatic was the decrease in R^2 from 19% to 6% (Table 1). As observed elsewhere [5], we found that the apparent performance and the optimism increased when smaller samples were considered. For example, the apparent ROC area was 0.74 (optimism 0.070) in samples of $n = 376$ and 0.73 (optimism 0.056) in the full sample of $n = 555$ (Table 1). Therefore, the assessment of expected optimism is especially important for small development samples.

The variability of model performance in independent validation data sets with 179 patients was large, which led to a limited power to detect true changes in model performance. Also, the empirical distribution of the estimated performance was not fully normal, especially for R^2, whereas a normal distribution was assumed for our power calculations. Furthermore, heteroscedasticity may be present (i.e., the SE varies with the mean performance). The power calculations should therefore be interpreted as approximate. In a specific validation study, the SE of any performance measure might be estimated nonparametrically by bootstrapping (i.e., by repeating the calculation of the performance measure in bootstrap samples) [4]. For some measures, such as the ROC area, variance estimates in the validation study may also be obtained analytically [24].

The evaluation of a model in independent validation data poses a methodologic dilemma: Should we perform a standard two-sample test, or should we consider a one-sample test? A two-sample test would be limited in power by the number of patients in the development and validation data sets. Even with an infinite number of patients in the validation data, the power would not reach 100% in a two-sample test, which is problematic for small differences in performance between development and validation setting. To justify a one-sample test, we might reason that physicians apply a predictive model for their patients, assuming that the regression coefficients and performance estimates are fixed and valid. Once data for a substantial number of patients have been gathered, a one-sample test might be appropriate to test the null hypothesis that the performance in their patients is not markedly different from the postulated performance (e.g., ROC area = 0.75). Furthermore, because we are usually interested only in a decrease in performance, a one-sided P value might be considered. We found, however, that the power of a validation study would be limited even when a one-sample test with one-sided P values was applied. Subjectively, we consider a decrease in ROC area >0.05, a decrease in $R^2 > 5\%$, and a shrinkage by more than 0.2 (e.g., 0.8 to 0.6) as clinically meaningful. Hence, larger sample sizes than previously used ($n = 179$, 45 events) [10] are advisable for validation studies. Exact recommendations are speculative at this time, and no guidelines exist to our knowledge.

Our findings imply that the focus of a validation study should not be on the statistical testing of differences in performance when a validation data set of limited size is available. Instead, we might aim for a parsimonious adjustment of a previously developed model to local circumstances. We might re-estimate the model intercept to obtain adequate calibration-in-the-large and adjust the predictor effects with the observed calibration slope to correct previously obtained regression coefficients [25]. Such adjustments are possible with a relatively small data set, whereas

more extensive model revisions, such as re-estimating the coefficient for each predictor, require larger data sets.

Our findings further support the interpretation of the disappointing external validity as reported in our previous study [10]. In that study, the apparent ROC area decreased from 0.83 to 0.76 according to internal validation with bootstrapping. Because bootstrapping proved reasonably accurate in the present analysis, the previous internally validated ROC area estimate of 0.76 was likely reasonably unbiased. It was therefore appropriate to compare the externally observed ROC area of 0.57 with this value [10]. Second, the asymptotic SE of the latter estimate (0.051) agreed with that found in our simulations (empirical SE 0.052). This supports the validity of the claim in our previous study of statistical significance of the drop in performance [10].

We conclude that the bootstrap is an adequate tool to estimate the optimism of a moderately complex model selection strategy in a small data set. Replaying the model selection gives much better estimates than a naive approach that considers only the finally selected model. However, we studied only one modeling strategy in one medical problem, and other strategies in other areas may lead to different conclusions. For example, when the selected model is rather stable, as indicated by low *P* values for the selected predictors and high *P* values for nonselected predictors, replaying the model selection strategy may not be necessary to estimate the optimism honestly. The variability of performance estimates in independent samples is a concern in validation studies, which therefore require substantial samples sizes.

Acknowledgments

This study was inspired by the comments of an anonymous reviewer regarding the sampling variability of external validation studies. We gratefully acknowledge the contributions of all medical students and clinicians involved in data collection, especially Dr. G. Derksen-Lubsen (Juliana Children's Hospital, The Hague, The Netherlands). This work was supported by a fellowship from the Royal Netherlands Academy of Arts and Sciences (EWS).

References

[1] Harrell FE Jr, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143–52.

[2] Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303–25.

[3] 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.

[4] Efron B, Tibshirani R. An introduction to the bootstrap. In: Monographs on statistics and applied probability. New York: Chapman and Hall; 1993. p. xvi, 436.

[5] Steyerberg EW, Harrell FE Jr, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.

[6] Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Stat Med 1989;8:771–83.

[7] Chatfield C. Model uncertainty, data mining and statistical inference. J Royal Stat Soc A 1995;158:419–66.

[8] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73.

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

[10] Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol, in press.

[11] Stone M. Cross-validatory choice and assessment of statistical predictions. J Royal Stat Soc B 1974;36:111–47.

[12] Nagelkerke N. A note on a general definition of the coefficient of determination. Biometrika 1991;78:691–2.

[13] Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991;10:1213–26.

[14] Copas JB. Regression, prediction and shrinkage. J Royal Stat Soc B 1983;45:311–54.

[15] Steyerberg EW, Eijkemans MJ, Harrell FE Jr, et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19:1059–79.

[16] Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 1997;92:548–60.

[17] del Barrio E, Cuesta-Albertos JA, Matran C. Asymptotic stability of the bootstrap sampled mean. Stoch Proc Appl 2002;97:289–306.

[18] Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. Biometrics 1997;53:603–18.

[19] Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52:935–42.

[20] Ye J. On measuring and correcting the effects of data mining and model selection. J Am Stat Assoc 1998;93:120–31.

[21] Laupacis A, Sekar N, Stiell IG. Clinical prediction rules: a review and suggested modifications of methodological standards. JAMA 1997;277:488–94.

[22] Steyerberg EW, Eijkemans MJ, Van Houwelingen JC, et al. Prognostic models based on literature and individual patient data in logistic regression analysis. Stat Med 2000;19:141–60.

[23] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.

[24] Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat Med 1994;13:499–508.

[25] Van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stat Med 2000;19:3401–15.