Internal Validation of Risk Models in Clustered Data: A Comparison of Bootstrap Schemes


* Correspondence to Prof. Dr. K. G. M. Moons, Julius Center for Health Sciences and Primary Care, t.a.v. K. G. M. Moons, UMC Utrecht, P.O. Box 85500, 3508 GA Utrecht, the Netherlands (e-mail: k.g.m.moons@umcutrecht.nl).

Initially submitted December 23, 2011; accepted for publication September 25, 2012.

Internal validity of a risk model can be studied efficiently with bootstrapping to assess possible optimism in model performance. Assumptions of the regular bootstrap are violated when the development data are clustered. We compared alternative resampling schemes in clustered data for the estimation of optimism in model performance. A simulation study was conducted to compare regular resampling on only the patient level with resampling on only the cluster level and with resampling sequentially on both the cluster and patient levels (2-step approach). Optimism for the concordance index and calibration slope was estimated. Resampling of only patients or only clusters showed accurate estimates of optimism in model performance. The 2-step approach overestimated the optimism in model performance. If the number of centers or intraclass correlation coefficient was high, resampling of clusters showed more accurate estimates than resampling of patients. The 3 bootstrap schemes also were applied to empirical data that were clustered. The results presented in this paper support the use of resampling on only the clusters for estimation of optimism in model performance when data are clustered.

Abbreviation: ICC, intraclass correlation coefficient.

Risk models are important tools for providing estimates of patient outcome. For example, models have been developed to predict the presence of deep venous thrombosis in primary care patients (1) and to estimate the prognosis for patients with traumatic brain injury (2). Model performance that is estimated in the data used to develop the model (development data) will be higher than estimates in new, similar patients, especially when the development data set is relatively small. This optimism in model performance can be efficiently estimated with bootstrapping (3). Bootstrapping replicates the process of study sample generation from an underlying source population by drawing samples with replacement from the study sample, with each the same size as the study sample (4).

The regular bootstrap scheme that resamples only patients does not replicate the process of study sample generation when data are clustered (5). In a multicenter study, for example, a sample of centers is obtained from the underlying source population, followed by inclusion of patients (6, 7). Bootstrapping also assumes exchangeability of the units (e.g., patients) that are resampled (4), meaning that a patient resampled from Center A should be exchangeable with a patient from Center B. This contradicts the clustered data characteristic, which implies that patients from Centers A and B can be different. Hence, regular bootstrapping might not be the best resampling scheme to assess internal validity in clustered data (8).

Alternative bootstrap schemes have been proposed to estimate nonparametric confidence intervals of regression coefficients in clustered data (9, 10). Those schemes include resampling clusters rather than patients (cluster bootstrapping) and consecutively resampling clusters and patients (2-step bootstrapping). The alternative bootstrap schemes need to be evaluated for the estimation of optimism in model performance to guide the choice of a bootstrap scheme for internal model validation in clustered data.
We conducted a simulation study to investigate whether alternative bootstrap schemes give more accurate estimates of optimism in model performance than the regular bootstrap. We subsequently applied all bootstrap schemes to empirical data in which surgical patients at risk for postoperative nausea and vomiting were clustered by treating anesthesiologist.

SIMULATION STUDY

To evaluate the alternative bootstrap schemes, we designed a simulation study (Figure 1). We generated a source population from which study samples were drawn. Risk models were developed in the study samples and internally validated with the different bootstrap schemes. The optimism estimated by internal validation was compared with the "true" optimism estimated by applying the developed models in the source population.

Source population

We generated a baseline source population that included 100 centers, which can be interpreted like any other clustering that results from observations made on multiple subjects within a group. The number of patients per center was Poisson distributed, with the mean varying per center according to the exponential function of a normal distribution (mean, 5.7; standard deviation, 0.3). This resulted in a source population of more than 30,000 patients, with a median of 284 (range, 117–761) patients per center. The dichotomous outcome $Y$ was related to 3 continuous and 3 dichotomous predictors, $X_1$–$X_6$. The regression coefficients of all these predictors were 1. Two nuisance predictors were added (1 dichotomous, 1 continuous) with regression coefficients equal to 0 to induce optimism in performance. The 4 continuous predictors were normally distributed, with a mean of 0 and standard deviations of 0.2, 0.4, 0.6, and 1. The incidences of the 4 dichotomous predictors were 0.2, 0.3, 0.3, and 0.4. The intercept was chosen such that the incidence of the outcome $Y$ (18%) and the corresponding number of events per variable ($n=11$) were relatively low in the study samples, inducing additional optimism in performance. To induce clustering, we generated a random cluster effect with mean of 0 and variance of 0.17. This corresponds to an intraclass correlation coefficient (ICC) of 5% (11). The random cluster effect was correlated with the first predictor $X_1$ ($\rho=0.2$). The linear predictor $lp$ was calculated for each...

Table 1. Parameter Values in the Baseline and Alternative Source Populations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Baseline Population</th>
<th>Alternative Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of centers</td>
<td>100</td>
<td>7,000$^b$</td>
</tr>
<tr>
<td>ICC, %</td>
<td>5</td>
<td>15 or 30</td>
</tr>
<tr>
<td>Data structure</td>
<td>Unbalanced data</td>
<td>Balanced data</td>
</tr>
<tr>
<td>Incidence of the outcome, %</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>No. of true (noise) predictors$^c$</td>
<td>6 (2)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Random slope variance</td>
<td>0</td>
<td>0.18</td>
</tr>
<tr>
<td>Correlation predictor and random intercept</td>
<td>0.2</td>
<td>0.0 or 0.4</td>
</tr>
<tr>
<td>Patient or center characteristic$^d$</td>
<td>All predictors as patient characteristics</td>
<td>$X_2$ and $X_7$ as center characteristics</td>
</tr>
</tbody>
</table>

Abbreviation: ICC, intraclass correlation coefficient.
$^a$ The ICC was based on the random intercept variance.
$^b$ The number of patients remained similar to the baseline source population ($n=29,925$).
$^c$ Beta of true predictor is 1. Beta of nuisance predictor is 0.
$^d$ In this specific situation, the variance of the fourth predictor was changed from 0.2 to 0.4. ICC = 10%; otherwise, there was no second-level variance in the full model.
patient and transformed to probabilities for the outcome with the formula \( P(Y) = \frac{1}{1 + \exp(-lp)} \). The outcome value \( Y \) (1 or 0) was then generated by comparing \( P(Y) \) with an independently generated variable \( u \), which had a uniform distribution from 0 to 1.

We also investigated the bootstrap schemes in alternative source populations. We varied in the baseline population a single parameter among the parameters described in Table 1. Parameters that varied among source populations were the ICC (12, 13), the numbers of “true” predictors and noise predictors, the incidence of the outcome, the data structure (balanced vs. unbalanced data) (11), the number of centers, random slope variance, and the correlation between predictor \( X_1 \) and the random intercept (Table 1). The ICCs were relatively low (5%–30%) but were based on what was found in clinical research practice and in previous simulation studies (12, 13). In the source population with a random slope, the predictor effect of \( X_1 \) varied among centers such that the random slope variance was 0.18. In total, we generated 11 different source populations.

**Study sample**

Study samples were drawn according to the practice of study sample generation in a multicenter setting (6, 7). We randomly drew 500 study samples from the source population in 2 stages. First we sampled 20 centers, and then we sampled a total of 500 patients from the selected centers (2-stage sampling). The number of patients per center differed and was representative of the distribution in the source population.

We also investigated the influence of alternative study sampling strategies. We varied the numbers of patients and centers in the study samples. Additionally, we conducted 1-stage sampling (Table 2) with either patients or centers as the only sampling unit. When centers were the sampling unit, all patients of the selected centers were included in the study sample.

**Model development and performance measures**

The risk models were fitted with mixed-effect logistic regression analysis and included all 8 predictors (6 true predictors, 2 noise predictors). Predicted risks were then compared with observed outcomes. The risk predictions were based on the estimated predictor effects and the model intercept and ignored the random cluster effect. The rationale for this approach was that the random cluster effect would be unknown for new centers (14).

The model performance was assessed with the concordance index (15) and the calibration slope (16, 17). The calibration slope was estimated with standard logistic regression analysis, modeling the outcome of interest as the dependent variable and the linear predictor as independent variable: \( \logit(y) = a + b \times lp \). In an ideal situation, the predicted and observed probabilities would be the same, and the value of the calibration slope \( b \) would be 1. Because standard performance measures ignore the clustered data structure, they can be considered as overall measures. To take clustering into account in the model evaluation, we assessed the predictive performance in individual centers (within-cluster performance). The within-cluster concordance index was estimated as the average of the concordance indexes of the centers (18). Within-cluster calibration was assessed with a mixed-effect logistic regression model, with random effects for the intercept and linear predictor (calibration slope).

**The bootstrap schemes and evaluation strategy**

The aim of the present study was to compare the regular bootstrap scheme (4) with 2 alternative bootstrap schemes for their ability to assess accurate estimates of optimism in performance of risk models derived from clustered data. The regular bootstrap procedure resampled patients with replacement (4). One alternative scheme resampled clusters, with replacement, and then included all patients of the selected clusters (cluster bootstrap). The other alternative bootstrap scheme first resampled clusters with replacement and then resampled patients from each selected cluster with replacement (2-step bootstrap). The numbers of patients and clusters varied among the 3 bootstrap procedures. In the regular bootstrap samples, the number of patients was equal to the number of patients from the study sample. In the cluster bootstrap and 2-step bootstrap samples, the number of clusters was equal to the number of clusters from the study sample. We used 200 bootstrap sampling repetitions to derive estimates of optimism. A model was developed in the bootstrap sample and evaluated in the bootstrap sample and the study sample. The difference in performance between the bootstrap sample and study sample is the bootstrap-estimated optimism. Furthermore, to estimate the true optimism, a model was developed in the study sample and evaluated in the study sample and in the source population. The difference in model performance between the study sample and the source population is the true optimism (Figure 1).

The whole process (sampling from source population to bootstrap validation) was repeated 500 times. To evaluate differences between the true and bootstrap-estimated optimism, we calculated the bias and the root mean squared
error. Bias was calculated as the average difference between bootstrap-estimated optimism and the true optimism in model performance. The mean squared error was calculated as ∑(bop_i − top_i)^2 / N, where bop_i and top_i were the bootstrap-estimated optimism and true optimism for sample i of 500 samples. Subsequently, we took the square root of the mean squared error to obtain the root mean squared error.

**Shrinkage factors**

Bootstrapping also is used to estimate the amount of required shrinkage likely to be present as a result of overfitting. The calibration slope estimated with the bootstrap procedure can be used to shrink regression coefficients of a risk model to avoid optimistic predictions in new patients. We compare the calibration slope with a slight modification of the heuristic shrinkage factor of van Houwelingen and le Cessie (19) as described by Copas (20). The heuristic shrinkage factor was calculated as (model \( \chi^2 - p \) / model \( \chi^2 \)). Here, the model \( \chi^2 \) is the likelihood ratio \( \chi^2 \) statistic for comparing a null model that contained a random intercept with the prediction model that also contained the 8 predictors. Hence, the \( p \) degrees of freedom related to this \( \chi^2 \) statistic are 8. The log-likelihoods used to calculate model \( \chi^2 \) statistics were derived from models fitted with the Laplace approximation (21, 22).

All analyses were performed in R, version 2.11.1 (23). We used the glmer function from the lme4 library to perform the multilevel analysis (24). The R script for the 2-step bootstrap is described in the Web Appendix.

**RESULTS OF SIMULATION STUDIES**

**Performance of the risk model**

Figure 2 shows the model performance for study samples with different ICCs (5% and 30%). The dashed horizontal line is the performance of a model developed and evaluated in the source population. The performance estimates shown in the boxplots were from models that were developed in 500 study samples. The performance estimates of these models were higher in the study samples (apparent performance) than in the source population (test performance). This indicated that model performance in the study samples was too optimistic. ICC values of 5% and 30% showed similar optimism in model performance.

**Evaluation of bootstrap schemes**

Table 3 shows the true optimism and the bias and variability of optimism in model performance as estimated by the different bootstrap schemes. The regular bootstrap scheme had bias similar to the cluster bootstrap when a total of 500 patients clustered in 20 centers were sampled from the baseline source population (Table 3). Optimism in the calibration slope within clusters was estimated most accurately with the cluster bootstrap. The 2-step bootstrap scheme overestimated the optimism. This scheme revealed some extreme model performance measures (e.g., overall calibration slopes < 0.2), particularly if small study samples were used (i.e., 300 patients). The number of bootstrap samples with extreme model performance measures was 467 of 100,000. Higher or lower numbers of patients in the study samples did not change the ranking of the accuracy of the bootstrap schemes (Web Table 1, A and B, available at http://aje.oxfordjournals.org/).

If the 500 sampled patients were clustered in many centers (\( n = 90 \)), optimism in the calibration slopes was estimated better by the cluster bootstrap than by the regular and 2-step bootstrap. Although the root mean squared error of optimism in the overall calibration slope was similar for the cluster and regular bootstraps (0.113 and 0.115, respectively), the bias was lower (0.0123 and -0.0249, respectively). Optimism in the concordance indexes was estimated accurately by the regular and cluster bootstrap schemes in data with 90 centers (Table 3). The 2-step bootstrap overestimated the optimism in all model performance measures. When 500 patients were clustered within 15 centers, the regular and cluster bootstraps had similar bias (Web Table 1C). However, study data containing 10 centers and 500 patients made resampling of clusters less efficient. Here, the regular bootstrap started to outperform the cluster bootstrap (Web Table 1D). The 2-step bootstrap overestimated optimism again (Web Table 1, C and D). One-stage sampling of 500 patients (1–100 centers) during study sample generation resulted in a lower bias for the cluster bootstrap than for the regular and 2-step bootstraps (Web Table 1E).

**Influence of variations on the source population**

We studied alternative values for ICC: 15% and 30%. The regular and cluster bootstraps showed similar bias and variability and gave more accurate estimates than the 2-step bootstrap in data with an ICC of 15% (Web Table 2A). For data with an ICC of 30%, we found that the cluster bootstrap led to smaller bias than the regular or 2-step bootstrap (Table 3). The bias in optimism estimated for the overall calibration slope was -0.0125 with the regular bootstrap, versus -0.0052 with the cluster bootstrap, when study samples contained 20 centers. These differences between the regular and cluster bootstrap schemes increased when 90 centers were included in the study sample or when the calibration slope was estimated within clusters.

One-stage sampling of centers was conducted from the alternative source population containing 7,000 centers (1–22 patients per center). We drew study samples of 100 centers from this source population. The ranking of quality of estimates in optimism by the bootstrap schemes was similar to the situation in which we sampled 90 centers from the baseline source population. The cluster bootstrap estimated optimism in the concordance index most accurately. The optimism in calibration slope was almost unbiased when estimated by the 2-step bootstrap (Web Table 2B).

We also investigated a model that included patient characteristics plus center characteristics (Table 1). We generated a source population with more variation between centers (ICC = 14%), which was explained partly by center characteristics. Each study sample included 50 centers to keep enough contrast in the dichotomous center characteristics. The cluster bootstrap schemes showed lowest bias in estimation of
Figure 2. Boxplots of apparent and test performance estimates in data with intraclass correlation coefficients (ICCs) of 5% and 30%. The risk models were developed in study samples of 500 patients clustered within 20 centers. The dashed horizontal line shows the performance of the model developed and evaluated in the source population. Internal Validation of Risk Models

optimism in this specific situation, as compared with the regular and 2-step bootstraps (Web Table 2H).

Other alternative situations that we considered were the data structure (balanced or unbalanced data), the incidence of the outcome, the number of noise predictors (4 instead of 2 in the baseline source population), higher random slope variances, and the correlation between predictor $X$ and the random effect. The regular and cluster bootstraps showed similar bias and variability and gave better estimates than the 2-step bootstrap (Web Table 2, C–G). However, in data in which predictor effects varied per center, the cluster bootstrap estimated optimism in the calibration slope more accurately than the regular and 2-step bootstraps (bias: $−0.0107$, $−0.0149$, and $0.0468$, respectively). For estimating optimism in the concordance index, the efficacies of the regular and cluster bootstraps were similar (Web Table 2I).

### Evaluation of shrinkage factors

Table 4 compares shrinkage factors estimated in study samples of several source populations. The first column (source) shows the true amount of shrinkage. The cluster bootstrap generally gave an accurate estimate of the amount of shrinkage, irrespective of how the calibration slope was estimated (overall calibration slope (Table 4) or within-cluster calibration slopes (data not shown)). The heuristic shrinkage factor slightly underestimated the amount of shrinkage.

### Table 4. Shrinkage Factors Estimated as Calibration Slopes With Different Bootstrap Procedures and the Heuristic Shrinkage Factor for Several Source Populations

<table>
<thead>
<tr>
<th>Source Population*</th>
<th>Sourceb</th>
<th>Regular Bootstrap</th>
<th>Cluster Bootstrap</th>
<th>Two-Step Bootstrap</th>
<th>Heuristic Shrinkage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.896</td>
<td>0.908</td>
<td>0.907</td>
<td>0.838</td>
<td>0.924</td>
</tr>
<tr>
<td>ICC, 30%</td>
<td>0.806</td>
<td>0.819</td>
<td>0.811</td>
<td>0.756</td>
<td>0.916</td>
</tr>
<tr>
<td>Incidence (Y), 10%</td>
<td>0.879</td>
<td>0.880</td>
<td>0.878</td>
<td>0.792</td>
<td>0.889</td>
</tr>
<tr>
<td>7,000 centers</td>
<td>0.877</td>
<td>0.915</td>
<td>0.899</td>
<td>0.873</td>
<td>0.916</td>
</tr>
</tbody>
</table>

Abbreviation: ICC, intraclass correlation coefficient.  
* See Table 2.  
b Values estimated in the source population indicate true amount of shrinkage.
APPLICATION TO THE PREDICTION OF POSTOPERATIVE NAUSEA AND VOMITING

As an illustrative example, we developed and internally validated a model for the risk of postoperative nausea and vomiting. We used data from adult surgical patients clustered by treating anesthesiologist at the University Medical Center Utrecht, Utrecht, the Netherlands (25). Patients were included in a randomized clinical trial between March 2006 and December 2007. Patients received antiemetic treatment either according to the risk as calculated by a risk model or according to regular care (reference group). We drew a small random sample of 250 patients from the reference group to induce optimism in model performance. The 250 patients were clustered by 32 anesthesiologists, with an ICC of 8.1%, indicating a positive dependency between outcomes of patients of the same anesthesiologist. This dependency could result from the use of a common treatment approach for patients treated by the same anesthesiologist. Predictive factors for postoperative nausea and vomiting included sex, age, history of postoperative nausea and vomiting or motion sickness, current smoking, abdominal or middle ear surgery, and the use of volatile anesthetics during surgery (26).

The estimated optimism by the different bootstrap schemes showed a pattern similar to the results of the simulation study. The cluster bootstrap showed the lowest estimates in optimism: 0.023 and 0.020 for optimism in overall and within-cluster concordance indexes, versus 0.028 and 0.045 for the regular bootstrap and 0.059 and 0.058 for the 2-step bootstrap. Estimates of optimism in the overall calibration slope were 0.13, 0.11, and 0.26 for the regular, cluster, and 2-step bootstraps, respectively. Estimates of optimism in within-cluster calibration slopes were 0.50, 0.28, and 0.67. We evaluated the model in the other patients of the empirical data set (n = 2,850) and calculated optimism in model performance by taking the difference in the performance in the sample and the performance in the reference group. The cluster bootstrap–estimated optimism was most similar to the optimism estimated with the reference group (Table 5).

DISCUSSION

We compared 3 different bootstrap schemes to internally validate risk models that were developed in clustered data. We showed that resampling of clusters gave the most accurate estimates of model optimism. We found that the regular bootstrap scheme, which resampled patients rather than clusters, could underestimate optimism if patients were clustered in a substantial (>20) number of clusters. The 2-step bootstrap, which first resampled centers with replacement and then resampled patients of each selected center, overestimated optimism.

Bootstrapping assumes exchangeability of the units being resampled (4, 9). In our simulated and empirical data, the dependency of patients was sometimes substantial (e.g., ICC = 30%), which implies that patients were more like other patients from the same center than patients from another center. Consequently, patients from the same center were not exchangeable with patients from another center. Therefore, the regular bootstrap, resampling patients, might be theoretically inappropriate in clustered (e.g., multicenter) data. Indeed, the estimates of optimism by the regular bootstrap had more bias than the cluster bootstrap in data with a high ICC. We evaluated the regular bootstrap for practical reasons: The cluster and 2-step bootstraps can be insensitive when the number of centers in the study sample is low. The bootstrap samples can contain even fewer centers and are more homogeneous than the study sample, which results in poor estimation of optimism. In this particular situation, the regular bootstrap might be preferred and can give accurate estimates of optimism.

In line with others, we found that the cluster bootstrap is preferable to the 2-step bootstrap. Xiao and Abrahamowicz (9) found that the 2-step bootstrap overestimates the variance for predictor effects at the individual level. Theoretically, the 2-step bootstrap scheme best resembles the real-life scenario in which a study sample contains a random set of clusters with randomly drawn patients. However, the 2-step bootstrap scheme introduced too much variability, as observed by Xiao and Abrahamowicz. All bootstrap strategies result in samples that differed from the original study sample. The 2-step bootstrap samples differed most from the study sample. The cluster and regular bootstrap will not be that different from the study sample. Hence, the models developed in the 2-step bootstrap samples showed relatively poor performance in the study sample, followed by a higher estimated optimism.

We examined different study samples to allow more generalizable conclusions. The number of centers was varied, because this influences the validity of the estimates of

Table 5. Optimism Estimated in the Empirical Data on Postoperative Nausea and Vomiting

<table>
<thead>
<tr>
<th></th>
<th>True Optimism</th>
<th>Regular Bootstrap</th>
<th>Cluster Bootstrap</th>
<th>Two-step Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall concordance</td>
<td>0.0159</td>
<td>0.0278</td>
<td>0.0233</td>
<td>0.0587</td>
</tr>
<tr>
<td>index</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concordance index</td>
<td>0.0262</td>
<td>0.0451</td>
<td>0.0196</td>
<td>0.0576</td>
</tr>
<tr>
<td>within anesthesiologist</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clusters</td>
<td>0.0763</td>
<td>0.1271</td>
<td>0.1116</td>
<td>0.2574</td>
</tr>
<tr>
<td>Overall calibration</td>
<td>0.1911</td>
<td>0.5007</td>
<td>0.2782</td>
<td>0.6671</td>
</tr>
<tr>
<td>slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>within anesthesiologist</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clusters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The model was developed in a small sample (n = 250) of the empirical data set. Optimism was assessed as the difference in model performance estimated in the sample and the rest of the empirical data set.
regression coefficients and random effects (12). The number of patients was varied, because this influences true optimism in model performance and the efficiency of bootstrapping as an internal validation technique (3, 27). Furthermore, we obtained study samples with 1- or 2-stage sampling. One-stage study sample generation, for instance, occurs when invited participants are from the general healthy population. Participants are invited randomly but can still be clustered by, for example, general practitioners. Two-stage sampling is common in multicenter trials, where first treatment centers or physicians are invited to participate, followed by patient inclusion. The cluster bootstrap had lower bias in study samples with many centers, generated by either 1- or 2-stage inclusion. The cluster bootstrap had lower bias in study samples with more than 50 patients with the event per estimated parameter (3, 31). In samples drawn from the source population, the number of events per variable was 11. Accordingly, models could be developed reasonably well, with relevant optimism in model performance.

We used the mean cluster effect when applying risk models for risk calculation (14). The rationale for this approach was that the cluster-specific effect was unknown for patients treated in new clusters that were not included in the study sample. Such a population-averaged estimate can be used to make a prediction for a patient in a new cluster, given the assumption that the new cluster is sampled from the same underlying population (14).

Shrinkage of regression coefficients may be applied to prevent predictions for new patients from being too extreme. A shrinkage factor can be estimated as the calibration slope in internal validation data or with the formula for a heuristic shrinkage factor. We used an adapted version of the heuristic shrinkage factor for multicenter data. Results showed that the heuristic shrinkage factor was still too weak. Further research is required to develop a heuristic shrinkage factor for clustered data (32, 33).

We conclude that the cluster bootstrap gives accurate estimates of optimism in the predictive performance of a risk model that is developed with clustered data. The cluster bootstrap resamples the data by drawing complete clusters with replacement. The regular bootstrap draws patients with replacement and, therefore, ignores the clustered nature of the data. Particularly in samples with a high ICC or with large numbers of clusters and small numbers of patients within clusters, the regular bootstrap will give biased results.

ACKNOWLEDGMENTS

Author affiliations: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands (W. Bouwmeester, K. G. M. Moons, M. J. C. Eijkemans, Y. Vergouwe); Department of Perioperative Care and Emergency Medicine, University Medical Center Utrecht, Utrecht, the Netherlands (T. H. Kappen, W. A. van Klei); and Department of Methodology and Applied Biostatistics, Institute of Health Sciences, VU University, Amsterdam, the Netherlands (J. W. R. Twisk).

This research was supported by the Netherlands Organization for Scientific Research (grant ZonMw 17088.25029 to W. Bouwmeester, K. G. M. Moons, and Y. Vergouwe; grants ZonMw 91208004 and ZonMw 918.10.615 to K. G. M. Moons).

Conflict of interest: none declared.

REFERENCES


