Research Article

Statistics in Medicine

Received: 5 May 2014, Accepted: 21 October 2016 Published online 28 November 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7179

A closed testing procedure to select an appropriate method for updating prediction models

Yvonne Vergouwe,[a*†] Daan Nieboer,[a] Rianne Oostenbrink,[b] Thomas P. A. Debray,[c] Gordon D. Murray,[d] Michael W. Kattan,[e] Hendrik Koffijberg,[c] Karel G. M. Moons[c] and Ewout W. Steyerberg[a]

Prediction models fitted with logistic regression often show poor performance when applied in populations other than the development population. Model updating may improve predictions. Previously suggested methods vary in their extensiveness of updating the model. We aim to define a strategy in selecting an appropriate update method that considers the balance between the amount of evidence for updating in the new patient sample and the danger of overfitting. We consider recalibration in the large (re-estimation of model intercept); recalibration (re-estimation of intercept and slope) and model revision (re-estimation of all coefficients) as update methods. We propose a closed testing procedure that allows the extensiveness of the updating to increase progressively from a minimum (the original model) to a maximum (a completely revised model). The procedure involves multiple testing with maintaining approximately the chosen type I error rate. We illustrate this approach with three clinical examples: patients with prostate cancer, traumatic brain injury and children presenting with fever. The need for updating the prostate cancer model was completely driven by a different model intercept in the update sample (adjustment: 2.58). Separate testing of model revision against the original model showed statistically significant results, but led to overfitting (calibration slope at internal validation = 0.86). The closed testing procedure selected recalibration in the large as update method, without overfitting. The advantage of the closed testing procedure was confirmed by the other two examples. We conclude that the proposed closed testing procedure may be useful in selecting appropriate update methods for previously developed prediction models. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: model updating; logistic regression; prediction model; closed testing procedure

Introduction

Clinical prediction models, developed with logistic regression, may require updating to new circumstances in other populations. The simplest update approach is to re-estimate all regression coefficients of the model (model revision). This approach however ignores previous evidence on the relative strength of predictor variables in the model. It may lead to poorer predictions in new patients than accepting the previously developed model if the sample size is limited [1]. Several more parsimonious update methods have been described, such as recalibration [1–4]. The model intercept can be adjusted to allow for a difference in baseline risk that is not reflected in the predictors in the model. A recalibration factor may be used that multiplies all regression coefficients with the same factor (calibration slope) to allow for a generally smaller or larger effect of all predictors. Only when effects of predictors have clearly changed, individual regression coefficients should be re-estimated ('model revision', Table I). In small samples, the

[a] *Center for Medical Decision Sciences, Department of Public Health, Erasmus MC, Rotterdam, the Netherlands*
[b] *Department of General Paediatrics, Erasmus MC-Sophia Children's Hospital, Rotterdam, the Netherlands*
[c] *Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, the Netherlands*
[d] *Centre of Population Health Sciences, University of Edinburgh, Edinburgh, UK*
[e] *Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH*
**Correspondence to: Y. Vergouwe, PhD, Department of Public Health, Erasmus MC, NA 2322, 3000 CA Rotterdam, The Netherlands.*
[†] *E-mail: y.vergouwe@erasmusmc.nl*

Copyright © 2016 John Wiley & Sons, Ltd.

Table I. Methods to update prediction models

Method	Label	Parameter	Number of parameters
0	Original model	None	0
1	Recalibration in the large	Intercept	1
2	Recalibration	Intercept and slope	2
3	Model revision	Re-estimate coefficients	p*

*p is equal to the number of regression coefficients (intercept not considered) in the original model.

1. when baseline risk is different

2. Update with a factor; multiply in all coefficient, might have smaller or larger effect on predictor

evidence on changed predictor effects should be strong before model revision is preferred over model recalibration.

Choosing the best updating method when limited sample size is available is not straightforward. Here, we consider three tests as a multiple testing problem, i.e. should we 1) update the model intercept; 2) recalibrate the model or 3) revise the model versus keeping the original model. These tests can be based on any performance measure of the prediction model in the dataset available to update the model, such as the *c*-statistic or R^2. We however use the likelihood ratio test statistic to test if an updated model provides a significantly better fit compared to the original model. Statistical testing for a change in other performance measures is conceptually equivalent to the likelihood ratio test, and it has previously been suggested to have inferior statistical properties [5].

If a family of hypotheses needs to be tested, a closed testing procedure can be used to maintain approximately the chosen type I error rate. The procedure includes an ordered sequence of test results and rejects hypotheses one at a time until no further rejections can be done [6,7]. An example of a procedure that is inspired by the closed testing procedure is the systematic search for the best fitting transformations of continuous variables with fractional polynomials [8]. The closed testing procedure is also the basis for several multiple testing procedures in genomics [9].

The multiple testing problem in the updating of prediction models combined with the clear ordering in extensiveness of the update methods, prompted us to develop a closed testing procedure. Hereto, the update methods are ordered from a prespecified minimum (keeping the original model as it was) to a prespecified maximum (model revision, i.e. re-estimation of all regression coefficients in the new data). The components of the test procedure are not new, but considering each separately would lead to an increase in Type I error. This is prevented by the proposed closed testing procedure. We demonstrate the procedure in three clinical examples: screening of patients for prostate cancer, predicting the prognosis of patients having traumatic brain injury and diagnosing children with fever for serious bacterial infections [10–12]. Properties of the proposed test procedure are also investigated in a simulation study.

Closed testing procedure for model updating

In this section, we propose the closed testing procedure for updating of a prediction model. We consider the situation that regression coefficients are available from a previously developed prediction model for a dichotomous outcome. When the model is applied in a new patient population, the question arises whether and to what extent the model needs to be updated for this patient population. We first briefly review the most frequently used update methods and how the updated regression coefficients can be estimated. We then define the sequentially rejective test procedure.

Update methods

Consider a linear predictor Z_0 that can be calculated as $\alpha + \sum \beta_i x_i$, where α is the model intercept, x represents the p predictor values in the new patient, and β represents the p original regression coefficients from the original prediction model. We review three previously described methods to update a logistic regression model (Table I) [1]. The first two methods are simple recalibration methods [4]; the third method includes re-estimation of all regression coefficients of the original model.

Method 1 considers only the model intercept and intends to correct 'calibration in the large'. The average predicted risks of the updated model become equal to the observed event rate in the update sample.

$$Z_1 = \alpha_{new} + Z_0$$

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hereto we fit a logistic regression model in a sample from the new population (update sample) with the intercept as the only free parameter and the linear predictor based on the original model (Z_0) as an offset variable (i.e. the slope is fixed at unity).

In method 2, we update both the model intercept α and the overall calibration slope β_{overall} by fitting a logistic regression model with Z_0 as the only covariate.

$$Z_2 = \alpha_{new} + \beta_{overall} Z_0$$

This method has also been labelled 'logistic calibration' [13].

Method 3 fits the regression coefficients of the original model anew and can be labelled 'model revision' [2]. Model revision does not consider variable selection or extension of the model.

$$Z_3 = \alpha_{new} + \sum_{i \in 1,\ldots,p} \beta_{new.i} x_i$$

The $\beta_{\text{new.i}}$ are the re-estimated coefficients for the p covariates as specified in the original model.

A closed testing procedure to select an update method

The aim of the closed testing procedure is to select an update method. The update methods (recalibration in the large, recalibration and revision) come with an increasing number of estimated parameters. Only if sufficiently strong evidence exists that individual regression coefficients are different in the update sample, the revision method is selected.

The procedure consists of a series of likelihood ratio tests of updated models against the original model. The procedure for a model with p regression coefficients (intercept not considered) is as follows:

A. Choose the nominal P value α for the hypothesis that the original model does not need updating.
B. Test the model revision (no. 3 in Table I) against the original model (no. 0 in Table I) at the α level using $p+1$ df. If the test is not significant, adopt the original model, otherwise continue.
C. Test the model revision against recalibration in the large (no. 1) at the α level using p df. If the test is not significant, adopt the updated model intercept, otherwise continue.
D. Test the model revision against the recalibrated model (no. 2) at the α level using p -1 df. If the test is not significant, adopt the recalibrated model, otherwise adopt the revised model. End of procedure.

The test at B assesses if any update of the original model is needed. The test at C examines the evidence for updating beyond calibration in the large. At D the choice is made between one overall adjustment for the regression coefficients (model recalibration) versus re-estimation of all individual regression coefficients. See Appendix for the R code.

Clinical examples

In three clinical examples, we updated prediction models in samples from other populations than the development population. We subsequently assessed the performance of the updated models. Calibration was assessed with the calibration intercept and slope [4]. Discrimination was assessed with the concordance index (c-index) [14,15]. We used Nagelkerke's R^2 as a measure of overall performance [16]. Particular in small samples, model performance estimates are too optimistic when model revision is used. The updated model describes the sample very well, but may calibrate and discriminate poorly in new patients. We used a bootstrap resampling procedure in the complete updating samples to correct for optimism if model revision was applied [17].

In order to study the behaviour of the closed testing procedure in smaller and larger samples, we simulated the situation of increasing sample sizes over time. This mimicked the situation of growing amount of evidence for the new population. We randomly draw patients without replacement from the available dataset and applied the closed test procedure to the first patients in the dataset, such that the event per variable ratio was equal to 5. Subsequently we increased the available sample size for updating the model such that the event per variable ratio increased by 1. We continued this until all available data were used when applying the closed testing procedure. At each step we recorded the update method selected by the closed testing procedure. We repeated this process 1,000 times by repeatedly drawing

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

patients without replacement from the available dataset. Subsequently we calculated the proportion of times each update method was selected at each event per variable ratio.

Example 1: Patients with prostate cancer

Data of 409 clinical patients, who were screened positive for prostate cancer, were previously used to develop a model for indolent cancer (versus important cancer). Patients were treated at Baylor College of Medicine or at the University Hospital Hamburg-Eppendorf [10]. Eighty patients (20%) had indolent cancer. The model that included 7 predictors was updated for a screening population using the data of the Rotterdam section of the European Randomized Study on Screening for Prostate Cancer (ERSPC) [18]. In total 278 men were screened and 136 of them (49%) had indolent cancer at radical prostatectomy.

Figure 1A shows the validity of the original model in the complete update sample (n = 278). The model predicted far too low probabilities of indolent cancer. Many more patients in the update sample had indolent cancer (49%) compared to the development set (20%). The corresponding calibration intercept for calibration-in-the-large was 2.58 (Table II; ideal value is 0.0). Updating the model intercept (update method 1) was hence required. The individual regression coefficients of the original model were on average correct as illustrated with the calibration slope of 1.01. Nevertheless, model revision showed differences in individual regression coefficients compared to the clinical model. These coefficients were overfitted (calibration slope corrected for optimism = 0.86). When model revision was used in the first 100 patients, overfitting was even more profound with a corrected calibration slope of 0.76 and optimism in c statistic of 0.03.

When the update methods were tested against the original model, all three methods showed statistically significant results for smaller and larger update samples (Figure 2A). The closed testing procedure selected 'recalibration in the large' in the complete sample. The test for model revision against the original model (step B) was statistically significant with $p < 0.001$ (Chi-square = 326, df = 8). The test for model revision against recalibration in the large (step C) was not statistically significant (Chi-square = 7.3, df = 7, p = 0.40), and 'recalibration in the large' was adopted as update method. Simulating the situation of increasing sample sizes over time showed that the closed testing procedure selected 'recalibration in the large' in nearly all the update samples (Figure 3A).

Example 2: Patients with traumatic brain injury

A model to predict six month mortality in traumatic brain injury (TBI) patients was previously developed with patient data from the North American and International Tirilazad trials (n = 2259) [11,19]. At six months after trauma, 517 patients (24%) had died. A model with three predictors (8 regression coefficients) [20] was updated using data of patients from a survey data set (European Brain Injury Consortium). This survey contained 822 patients; 281 patients died within 6 months after trauma (34%).

In this example, the difference in overall risk between development and update samples (24% versus 34%) did not influence calibration in the large with the calibration intercept close to 0.0 (-0.01, Table III). The calibration slope was larger than 1 (1.25, Figure 1B), which implies that the overall

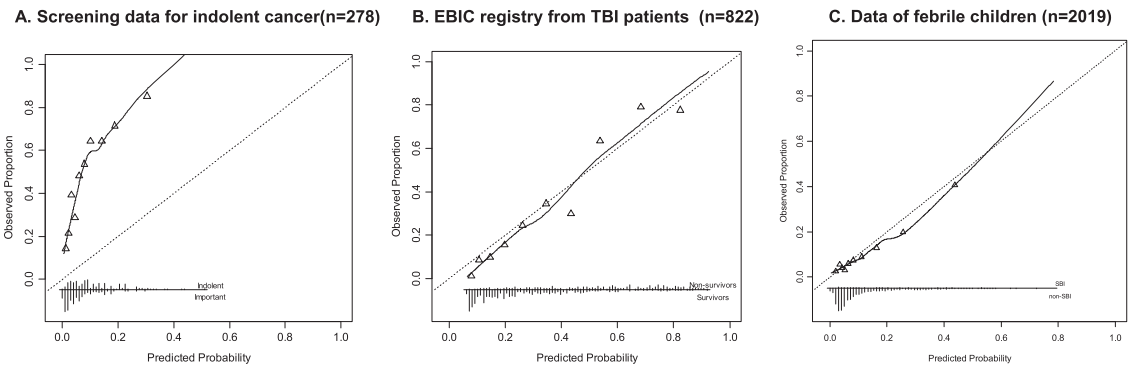

Figure 1. Validation plots of the original models applied in the update samples for indolent cancer (A), 6 month outcome in TBI patients (B) and SBI in febrile children (C). The triangles represent deciles of subjects grouped by similar predicted risk. The distribution of subjects is indicated with spikes at the bottom of the graph, separately for persons with and without the outcome

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table II. Logistic regression coefficients in the original clinical model for indolent cancer and the updated models for the screening setting (n = 278). Model performance in the screening setting is also shown.

	Clinical model	Re- calibration in the large	Re-calibration	Model revision
Coefficients				
PSA level, ng/ml [#]	- 1.09	- 1.09	- 1.10	- 1.44
Clinical stage T2a	0.17	0.17	0.17	0.08
Primary biopsy Gl.	- 0.30	- 0.30	- 0.30	0.37
Secondary biopsy Gl.	- 0.05	- 0.05	- 0.05	- 1.74
US prostate volume, 10 cc	0.20	0.20	0.20	0.32
Cancerous tissue, mm [#]	- 0.61	- 0.61	- 0.61	- 0.64
Positive cores, 10%	- 0.39	- 0.39	- 0.39	- 0.21
Intercept	1.11	3.67	3.70	6.37
Model performance				
Calibration intercept	2.58	0.00	0.00	- 0.01
Calibration slope	1.01	1.01	1.00	0.86*
c statistic	0.75	0.75	0.75	0.75*
R^2	0.25	0.25	0.25	0.21*

Gl., Gleason score; US, ultrasound
[#]log transformed
*model performance, corrected for optimism with bootstrapping

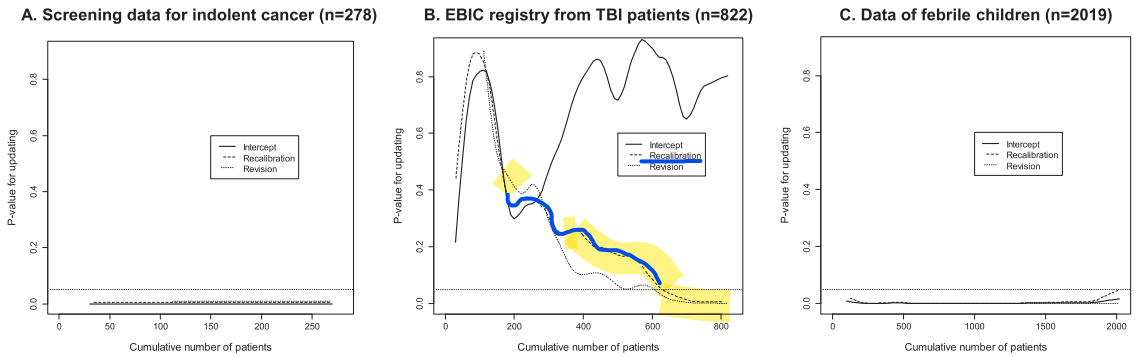

Figure 2. Test results for different methods to update the prediction models. Update methods were tested against the original model with increasing sample sizes over time.

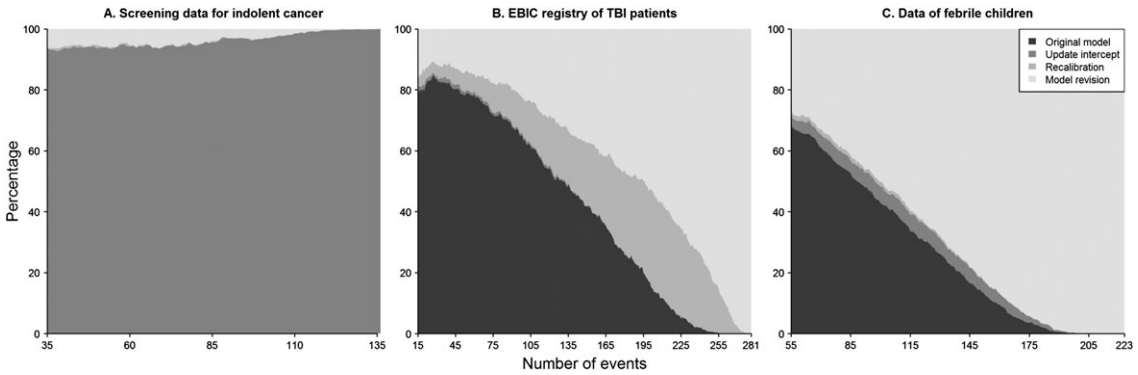

Figure 3. Selected update methods with the closed testing procedure for increasing sample sizes.

predictive effects in the original model were not extreme enough for the patients included in the survey. Low predicted risks corresponded to lower observed frequencies of death; higher risks to higher frequencies. Particularly the predictive effects for age and pupil reactivity (both sides) tended to be larger (Table III). Model revision slightly increased the discriminative ability of the model (c statistic corrected for optimism = 0.85 versus 0.84 for the original model). When model revision was applied in the first 200 patients, we found clear optimism (a corrected calibration slope of 0.82 and optimism in c statistic of 0.02).

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table III. Logistic regression coefficients in the original TBI model (Tirilazad) and the updated models for the EBIC registry (n = 822). Model performance in the EBIC registry is also shown.

	Tirilazad	Re- calibration in the large	Re-calibration	Model revision
Coefficients				
Age, 10 years	0.28	0.28	0.35	0.42
Motor score				
Extension	-0.48	-0.48	-0.60	0.07
Abnormal flexion	-0.87	-0.87	-1.09	-0.80
Normal flexion	-1.31	-1.31	-1.64	-1.23
Localises	-1.81	-1.81	-2.27	-1.78
Obeys command	-1.92	-1.92	-2.41	-1.35
Pupil reaction negative One side	0.56	0.56	0.70	0.64
Both sides	0.97	0.97	1.22	1.63
Intercept	-1.19	-1.19	-1.36	-2.22
Model performance				
Calibration intercept	-0.01	0.00	0.00	-0.01
Calibration slope	1.25	1.25	1.00	0.96*
c statistic	0.84	0.84	0.84	0.85*
R^2	0.41	0.41	0.42	0.41*

*model performance, corrected for optimism with bootstrapping

Separate testing of the updated models against the original model in smaller and larger samples showed that updating the model intercept was not needed (Figure 2B). In update samples of around 660 patients or more the tests for recalibration were statistically significant. Tests for model revision became statistically significant after inclusion of 620 patients. In the complete update sample, the closed testing procedure selected model revision. All three tests were statistically significant with p = 0.002 (step B: model revision against the original model, Chi-square = 27, df = 9), p = 0.001 (step C: model revision against recalibration in the large, Chi-square = 27, df = 8), and p = 0.013 (step D: model revision against the recalibrated model Chi-square = 18, df = 7). The closed testing procedure selected recalibration or the original model as the preferred method in most of the small simulated update samples (Figure 3B). As sample size increased model revision became the preferred update method in almost all simulated update samples.

Example 3: Children presenting with fever

A prediction model for serious bacterial infection (SBI) was previously developed with data from 1750 children with fever presenting at the emergency department of the Sophia's Children hospital, Rotterdam, the Netherlands (222, 13% had SBI). The model that included 11 predictors (12 regression coefficients) was updated with new data collected in the same hospital. In total, 2019 children were analysed with 223 (11%) patients having SBI.

The original model showed good performance in more recently treated febrile children with some over prediction (Figure 1C). Calibration intercept and slope were slightly different from the ideal values 0 and 1 (- 0.18 and 0.96, Table IV). Model revision slightly increased the discriminative ability of the model (c statistic corrected for optimism = 0.77 versus 0.76 for the original model).

When the update methods were tested against the original model, all three methods showed statistically significant results for smaller and larger update samples (Figure 2C). The closed testing procedure selected 'model revision' in the complete sample. All three tests were statistically significant with p = 0.001 for all three (step B: Chi-square = 36, df = 13; step C: Chi-square = 34, df = 12; step D: Chi-square = 33, df = 11). For smaller update samples the original model was chosen in 65% of the simulations and model revision in 30% (Figure 3C). As sample size increased model revision was chosen as the preferred updating method in the majority of simulated update samples.

Simulation study

To investigate the type 1 error rate of the closed testing procedure we performed a simulation study based on the three clinical examples. For each example, patient predictor data were drawn with replacement from the original dataset. The binary outcome value for each patient was generated by comparing

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table IV. Logistic regression coefficients in the original SBI model and the updated models for the new patients (n = 2019). Model performance in the new patients is also shown.

	Original	Re- calibration in the large	Re-calibration	Model revision
Coefficients				
Age, years (1 m-1 yr)[†]	-0.65	-0.65	-0.62	-0.66
Age, years (>1 yr) [†]	0.07	0.07	0.07	0.07
Girl	0.34	0.34	0.33	0.27
Temperature, °C[††]	0.13	0.13	0.12	0.13
Duration fever, days	0.10	0.10	0.10	0.16
Tachypnoea	0.15	0.15	0.14	0.33
Tachycardia	-0.05	-0.05	-0.05	0.48
Oxygen saturation, <94%	1.03	1.03	0.99	0.89
Capillary refill time, >3 s	0.28	0.28	0.27	0.34
Chest wall retractions	-0.04	-0.04	-0.04	0.42
Ill appearance	0.32	0.32	0.31	0.07
CRP, mg/l[#]	0.84	0.84	0.81	0.78
Intercept	-5.42	-5.60	-5.45	-5.77
Model performance				
Calibration intercept	-0.18	0.0	0.0	-0.09
Calibration slope	0.96	0.96	1.0	0.94*
c statistic	0.76	0.76	0.76	0.77*
R^2	0.19	0.19	0.19	0.19*

[†]Age was added using a piecewise linear term which can be calculated as -.65*min(1,Age) + 0.07*(Age-1)$_+$. Where (x) $_+$ = max(x,0).

[††]centred at 35 degrees;

[#]log transformed

*model performance, corrected for optimism with bootstrapping

For 7 coefficients in the model and EPV = 5, you need 7×5=35 events. With outcome incidence = 50%, the sample size must include 35×2=70 (half with the outcome, half without).

the predicted probability based on the fitted model on the full sample (model revision) with an independently generated variable u_i having a uniform distribution from 0 to 1, with $Y_i = 1$ if $p_i \geq u_i$ and 0 otherwise.

We considered four event per variable ratios for each clinical example, i.e. epv = 5, 10, 15 or 20. Given the epv, outcome incidence and number of regression coefficients in the model samples included 70, 140, 210 and 280 patients for the indolent prostate cancer example (7 coefficients), 118, 236, 354 and 472 patients for the TBI example (8 coefficients), and 545, 1090, 1635 and 2180 patients for the SBI example (12 coefficients). For each scenario we generated 2000 samples and applied the closed testing procedure. The proportion of test results indicating that the model should be updated equalled the type I error rate.

When small update samples were available the type I error rate was slightly above the nominal 0.05 level (Table V). As sample size increased the observed type I error rate decreased to the 0.05 level.

We measured in the same simulation set up the average mean squared error (MSE) of the estimated regression coefficients. Updating the model according to the closed testing procedure showed lower average MSE than model revision (Table VI). When the regression coefficients were shrunk, the average MSE was lower with values of 0.4 and 0.01 for the model revision and closed testing procedure in the indolent example; 0.13 and 0.01 in the TBI example and 0.02 and 0.0003 in the SBI example at event per variable ratio of 5.

Table V. Type I error rate for the closed testing procedure at different event per variable ratios.

Event per variable ratio:	5	10	15	20
Model				
Indolent	0.08	0.07	0.06	0.06
TBI	0.07	0.06	0.06	0.06
SBI	0.07	0.06	0.05	0.05

Indolent: indolent cancer; TBI: traumatic brain injury; SBI: severe bacterial infection

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table VI. Average mean squared error of the estimated coefficients with model revision or the closed testing procedure at different event per variable ratios.

Event per variable ratio	5		10		15		20	
	Revision	Closed	Revision	Closed	Revision	Closed	Revision	Closed
Model								
Indolent	8.0	1.0	3.0	0.35	1.2	0.17	0.58	0.1
TBI	1.5	0.31	0.36	0.08	0.2	0.04	0.14	0.03
SBI	1.3	0.04	0.10	0.01	0.06	0.005	0.05	0.004

Indolent: indolent cancer; TBI: traumatic brain injury; SBI: severe bacterial infection.
Revision: re-estimation of all regression coefficients; Closed: model was based on the closed testing procedure.

Discussion

Studying the performance of a prediction model in new patient data is a valuable step before using the model in clinical practice. When models are applied in populations with local or contemporary circumstances that differ from the development population, the model may need adjustments. Data collected in the new population may then be used to update the model to perform well given the new circumstances [21].

We proposed a closed testing procedure to select an appropriate method for updating a previously developed prediction model. Application of the procedure in three clinical examples showed that more parsimonious methods such as recalibration in the large or recalibration were selected, if the amount of evidence for model revision was relatively small. The closed testing procedure can hence warn the researcher that re-estimation of all regression coefficients from a prediction model for the own situation may result in an overfitted model.

The available sample size in the three clinical examples might be considered sufficient to apply model revision instead of updating the original prediction model. However previous studies have shown that if the development sample is relatively large compared to the update sample it is still advantageous to incorporate the previous evidence of the relative predictor strengths and update the original prediction model rather than re-estimation of all predictor effects [1].

When the closed testing procedure selects model revision as update method, the updating sample may still be relatively small and overfitting can occur. Some kind of shrinkage of the regression coefficients should be considered [22]. We did not consider the shrinkage step in the closed testing procedure itself, since it would be difficult to assess the correct number of degrees of freedom [23]. Further, the number of degrees of freedom can become below 2 with penalized maximum likelihood estimation, with alpha not equal to the chosen value for the closed testing procedure [24].

We considered here relatively small update samples and limited therefore the update methods; model extension in which extra predictors are included in the prediction model was not included in the closed testing procedure. Further, the closed testing procedure is a frequentist approach to combine prior knowledge incorporated in the established prediction model with new data of the updating sample. Bayesian modelling would be an alternative update method [25]. Further research should elaborate on these issues.

We assume that the used update sample is a representative and random sample of the considered population. The updated model results from new regression analyses with limited sample size. As a consequence, evidence on reproducibility of the updated model for the considered populations is relevant [26,27]. This internal validity may be studied with bootstrap resampling or with new patients from the same population.

Selecting an update method with the closed testing procedure is particularly relevant if an established prediction model is available but not the individual patient data used to develop that model. When the development data are also available, meta-analytical techniques can be applied to investigate the heterogeneity in predictor effects and to adjust the baseline risk if necessary [28,29]. If the observed heterogeneity between both samples is small, development and update samples can be combined to develop a model for both the development and update sample. If the primary interest is to develop a prediction model that is valid in the population of the update sample, adapting the model using the closed testing procedure is sufficient.

In conclusion, we proposed and illustrated a closed testing procedure to select methods for updating a previously developed prediction model. The closed testing procedure selects parsimonious update methods when sample sizes are relatively small. Only if strong evidence is present in the new data that individual regression coefficients should be re-estimated, model revision is selected.

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Appendix. R code for closed testing procedure

```
ClosedTest <- function(coefs, X, y){
# Implement closed testing procedure (Version: 11-01-2013)
# Arguments:
#   coefs: Vector containing the regression coefficients of the model that
#          is updated.
#   X: predictor matrix
#   y: outcome vector
# Results:
#   coef_new: regression coefficients of chosen model
require(rms)
if(class(X)=="data.frame"){
  X <- data.matrix(X)
}
if(ncol(X)!=(length(coefs)-1)){
  stop("Number of predictors not equal to the number of coefficients")
}
n_coefs <- length(coefs)
lp_old  <- X %*% as.matrix(coefs[2:n_coefs])
# Calculate updated model intercept
intercept <- lrm.fit(y = y, offset = lp_old)$coefficients
coefs_int <- c(intercept , coefs[2:n_coefs])
# Calculate coefficients after recalibration
recal       <- lrm.fit(x = lp_old, y = y)$coefficients
coefs_recal <- c(recal[1], recal[2] * coefs[2:n_coefs])
# Calculate coefficients after model revision
coefs_refit <- lrm.fit(x = X, y = y)$coefficients
# Calculate the log-likelihood of the different models
lp              <- cbind(1, X) %*% coefs
ll_original     <- sum(y * lp - log(1 + exp(lp)))
lp              <- cbind(1, X) %*% coefs_int
ll_intercept    <- sum(y * lp - log(1 + exp(lp)))
lp              <- cbind(1, X) %*% coefs_recal
ll_recalibration <- sum(y * lp - log(1 + exp(lp)))
lp              <- cbind(1, X) %*% coefs_refit
ll_revision     <- sum(y * lp - log(1 + exp(lp)))
# Calculate difference in log-likelihood for testing of the models
dev_original      <- -2 * ll_original + 2 * ll_revision
dev_intercept     <- -2 * ll_intercept + 2 * ll_revision
dev_recalibration <- -2 * ll_recalibration + 2 * ll_revision
# See if difference in model fit was significant
test1 <- (1 - pchisq(dev_original, ncol(X) + 1)) < 0.05
test2 <- (1 - pchisq(dev_intercept, ncol(X))) < 0.05
test3 <- (1 - pchisq(dev_recalibration, ncol(X) - 1)) < 0.05
# See which model is chosen, index_test indicates the chosen model
# 1. Original model
# 2. Model with updated intercept
# 3. Recalibrated model
# 4. Revised model
test_original      <- 1 * (!test1)
test_intercept     <- 2 * ((test1)&(!test2))
test_recalibration <- 3 * ((test1)&(test2)&(!test3))
test_revision      <- 4 * ((test1)&(test2)&(test3))
index_test <- (test_original + test_intercept + test_recalibration +
               test_revision)
coefs_result <- rbind(coefs, coefs_int, coefs_recal, coefs_refit)
```

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

```
# Output of the function
new_coefs <- coefs_result[index_test, ]
model     <- c("Original Model", "Model with updated intercept",
               "Recalibrated model", "Model Revision")[index_test]
cat("Method chosen by closed test procedure:\n", model, "\n",
    "Resulting coefficients:\n", new_coefs, "\n")
res <- list(model = model, coefs = coefs_result)
return(res).
```

Financial Support

The Netherlands Organization for Scientific Research (ZonMw 1709.925.039 and 9120.8004)

References

1. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine* 2004; **23**:2567–2586.
2. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 2000; **19**:3401–3415.
3. van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine* 1995; **14**:1999–2008.
4. Cox DR. Two Further Applications of a Model for Binary Regression. *Biometrika* 1958; **45**:562–565.
5. Vickers A, Cronin A, Begg C. One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology* 2011; **11**:13.
6. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
7. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Statist* 1979; **6**:65–70.
8. Ambler G, Royston P. Fractional polynomial model selection procedures: investigation of type I error rate. *Journal of Statistical Computation and Simulation* 2001; **69**:89–108.
9. Goeman J, Solari A. Multiple hypothesis testing in genomics. *Statistics in Medicine* 2014; **33**:1946–1978.
10. Kattan MW, Eastham JA, Wheeler TM, Maru N, Scardino PT, Erbersdobler A, Graefen M, Huland H, Koh H, Shariat SF, Slawin KM, Ohori M. Counseling Men With Prostate Cancer: A Nomogram for Predicting the Presence of Small, Moderately Differentiated, Confined Tumors. *The Journal of Urology* 2003; **170**:1792–1797.
11. Hukkelhoven CWPM, Steyerberg EW, Farace E, Habbema JDF, Marshall LF, Maas AIR. Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. *Journal of Neurosurgery* 2002; **97**:549–557.
12. Oostenbrink R, Moons KGM, Donders ART, Grobbee DE, Moll HA. Prediction of bacterial meningitis in children with meningeal signs: reduction of lumbar punctures. *Acta Paediatrica* 2001; **90**:611–617.
13. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer: New York, 2001.
14. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
15. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Jama* 1982; **247**:2543–2546.
16. Nagelkerke NJD. A note on the general definition of the coefficient of determination. *Biometrika* 1991; **78**:691–692.
17. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology* 2001; **54**:774–781.
18. Roobol MJ, Kirkels WJ, Schroder FH. Features and preliminary results of the Dutch centre of the ERSPC (Rotterdam, the Netherlands). *BJU International* 2003; **92**(Suppl 2):48–54.
19. Marshall LF, Maas AIR, Marshall SB, Bricolo A, Fearnside M, Iannotti F, Klauber MR, Lagarrigue J, Lobato R, Persson L, Pickard JD, Piek J, Servadei F, Wellis GN, Morris GF, Means ED, Musch B. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. *Journal of Neurosurgery* 1998; **89**:519–525.
20. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Medicine* 2008; **5**:1251–1260.
21. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. *Medical Decision Making* 1993; **13**:49–58.
22. 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. *Statistics in Medicine* 2004; **23**:2567–2586.
23. Stein C. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* 1981; **9**:1135–1151.
24. Zou H, Hastie T, Tibshirani R. On the "degrees of freedom" of the LASSO. *Ann. Statist.* 2007; **35**:2173–2192.
25. Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Statistics in Medicine* 2012; **31**:2697–2712.

Copyright © 2016 John Wiley & Sons, Ltd.

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

26. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999; **130**:515–524.
27. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology* 2015; **68**:279–289.
28. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine* 2013; **32**:3158–3180.
29. Abo-Zaid G, Guo B, Deeks JJ, Debray TPA, Steyerberg EW, Moons KGM, Riley RD. Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology* 2013; **66**:865–873. e4

10970258, 2017, 28, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.7179 by University Of British Columbia, Wiley Online Library on [15/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Copyright © 2016 John Wiley & Sons, Ltd.