STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:2567–2586 (DOI: 10.1002/sim.1844)

Validation and updating of predictive logistic regression models: a study on sample size and shrinkage

Ewout W. Steyerberg[1,*,†], Gerard J. J. M. Borsboom[1], Hans C. van Houwelingen[2], Marinus J. C. Eijkemans[1] and J. Dik F. Habbema[1]

[1]*Center for Clinical Decision Sciences, Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands*
[2]*Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands*

SUMMARY

A logistic regression model may be used to provide predictions of outcome for individual patients at another centre than where the model was developed. When empirical data are available from this centre, the validity of predictions can be assessed by comparing observed outcomes and predicted probabilities. Subsequently, the model may be updated to improve predictions for future patients.

As an example, we analysed 30-day mortality after acute myocardial infarction in a large data set (GUSTO-I, $n = 40\,830$). We validated and updated a previously published model from another study (TIMI-II, $n = 3339$) in validation samples ranging from small (200 patients, 14 deaths) to large (10 000 patients, 700 deaths). Updated models were tested on independent patients. Updating methods included re-calibration (re-estimation of the intercept or slope of the linear predictor) and more structural model revisions (re-estimation of some or all regression coefficients, model extension with more predictors). We applied heuristic shrinkage approaches in the model revision methods, such that regression coefficients were shrunken towards their re-calibrated values. Parsimonious updating methods were found preferable to more extensive model revisions, which should only be attempted with relatively large validation samples in combination with shrinkage. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: logistic regression; validation; updating; shrinkage

1. INTRODUCTION

Logistic regression may well be used to develop predictive models for dichotomous outcomes, such as short-term mortality [1]. When a previously developed model is applied in another centre, and/or in a more recent time period, the external validity (or generalizability) of model predictions is important [2]. When empirical data are available, the external validity can be assessed [3]. Also, we may consider updating of the previously

*Correspondence to: Ewout W. Steyerberg, Center for Clinical Decision Sciences, Ee2093, Department of Public Health, Erasmus Medical Center, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.
†E-mail: e.steyerberg@erasmusmc.nl

Received December 2001
Accepted January 2004

developed model [4, 5], such that the predictive model is adjusted to local and/or contemporary circumstances.

We may consider various performance measures in the assessment of external validity, including calibration (agreement between predicted probabilities and observed frequencies) and discrimination (ability to distinguish favourable from unfavourable outcomes) [1, 2]. Calibration may conveniently be studied in the context of a general calibration model, where the linear predictor based on the previously developed model is the only covariate [6]. This model has only two free parameters: intercept α and (calibration) slope β. A simple updating method might focus on re-calibration, i.e. that the updated model has a new intercept α and new regression coefficients based on multiplication of the original coefficients with β. An even simpler updating method might only adjust the intercept α, assuming a β of unity. These approaches have been followed for updating of a previously developed model in the context of risk-adjustment [7, 8] and prediction [5, 9]. We may also consider more extensive updating methods ('model revision'), such as re-estimation of regression coefficients of some or all predictor variables, and considering more covariables for inclusion of the model ('model extension') [5, 8].

Re-calibration methods are attractive because of their stability, which is related to the fact that few parameters are estimated [5]. Their disadvantage is a potential for bias in the individual regression coefficients. In contrast, model revision is expected to lead to a lower bias but higher variance in the updated model, since more parameters are estimated. Therefore the sample sizes of both the validation data set and the development data set are crucial in the choice of updating method. We aimed to study the influence of sample size on the performance of alternative updating methods.

Further, 'shrinkage' methods can be useful when a predictive model is estimated in a small data set with relatively many parameters [1, 10–14]. Traditionally, regression coefficients are shrunken towards zero, which is in the spirit of empirical Bayes analysis with a non-informative prior [15]. Here, we consider shrinkage of regression coefficients of revised models towards their re-calibrated values. A motivation for this approach is given in Section 2. Our second aim was to determine the benefit of shrinkage methods when updating a previously developed model.

As an example, we apply re-calibration and model revision methods to logistic regression models in a case study of 30-day mortality in patients with an acute myocardial infarction (GUSTO-I, $n = 40\,830$). Validation and test data sets of different sizes were randomly drawn from geographical regions within this large data set (Section 3). The updating methods are presented in Section 4, followed by a description of the performance measures that we consider in Section 5. We present results from some exemplary geographical subsets in Section 6, and results from more systematic simulation studies in Section 7. We discuss our findings in Section 8.

2. GENERALIZED SHRINKAGE AND (SHRUNKEN) RE-CALIBRATION

We will first discuss the simple linear regression model and generalize our findings to generalized linear models (GLMs) such as logistic regression at the end of the section.

2.1. Generalized shrinkage

We consider the linear regression model M given by $Y = \alpha_{\text{model}} + \beta_1 X_1 + \cdots + \beta_p X_p + e$, with Y a continuous outcome variable, α_{model} the intercept in the model, β_1–β_p regression coefficients, X_1–X_p covariates, and e an error term with mean 0 and variance σ^2. Estimates of the regression coefficients $\hat{\beta}_i$ are readily obtained by ordinary least square estimation in a training set containing n subjects. It has been shown that shrinkage procedures are useful to decrease the expected mean square error for future observations [10, 11, 16] if the number of covariates is relatively large (at least 3). The usual shrinkage model that shrinks towards the overall mean is given by $Y_{\text{pred}} = \bar{Y} + \hat{c}\hat{\beta}_1(X_1 - \bar{X}_1) + \cdots + \hat{c}\hat{\beta}_p(X_p - \bar{X}_p)$. The amount of shrinkage needed can be derived by bootstrapping or cross-validation. It is also possible to estimate the shrinkage factor from the model fit [10, 11]. A so-called heuristic shrinkage factor c has been proposed as $\hat{c} = (F_{\text{model}} - 1)^+ / F_{\text{model}}$, where the F statistic is calculated from the least square estimation. We truncate c at zero, such that c is in the interval [0,1].

The shrunken model can also be written as a compromise model: $Y_{\text{pred}} = (1-\hat{c})\bar{Y} + \hat{c}(\hat{\alpha}_{\text{model}} + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p) = (1 - \hat{c})\bar{Y} + \hat{c}\hat{Y}$, which is a weighted average of the null-model and the fitted model.

Shrinkage towards the mean can easily be generalized to shrinkage towards a more general null-model. Let M_0 be any linear submodel (including the constant term) of dimension $p_0 + 1$. It might consist of a selection of p_0 covariates + the constant, but it could as well be any other linear submodel. Let $F_{1|0}$ be the F-test statistic for testing the general model M versus the submodel M_0. A model that shrinks towards this null-model can simply be formulated as $\hat{Y}_{\text{pred}} = \hat{Y}_0 + \hat{c}(\hat{Y}_1 - \hat{Y}_0) = (1 - \hat{c})Y_0 - \hat{c}\hat{Y}_1$. Using the same arguments as in the case of simple shrinkage it can be shown that a heuristic shrinkage estimator is given by $\hat{c} = (F_{1|0} - 1)^+ / F_{1|0}$. Again, shrinkage can be expected to reduce the mean squared prediction error if the difference in dimension $p - p_0$ is relatively large.

The generalized shrinkage can be useful if the covariates can be split in one group of a few well-established covariates and another group of many less important covariates. Another application is the setting of re-calibration, as we will discuss below.

2.2. Re-calibration

Suppose we have a validation sample of m new subjects, drawn as a random sample from a similar population as the training set. Predictions for Y are calculated as $\tilde{Y}_{\text{pred}} = \tilde{\alpha}_{\text{model}} + \tilde{\beta}_1 X_1 + \cdots + \tilde{\beta}_p X_p$. The validity of this model can be tested at different levels. The re-calibration approach is to define a linear predictor Z as $Z = \tilde{\alpha}_{\text{model}} + \tilde{\beta}_1 X_1 + \cdots + \tilde{\beta}_p X_p$, and to use this linear predictor in a regression model: $\hat{Y}_{\text{cal}} = \hat{\alpha}_{\text{overall}} + \hat{\beta}_{\text{overall}} Z$. In the ideal case of perfect validity, $\hat{\alpha}_{\text{overall}} = 0$ and $\hat{\beta}_{\text{overall}} = 1$. These parameters can be tested with ANOVA or Wald statistics. If α and/or β significantly deviate from the ideal case, there is a need to recalibrate the model. Re-calibrated regression coefficients can be calculated as $\hat{\beta}_{\text{cal}} = \hat{\beta}_{\text{overall}} \tilde{\beta}_i$. One could say that the re-calibrated model borrows the relative effects (ratios) of the regression coefficients from the model in the training set.

We can compare the calibration model $\hat{Y}_{\text{cal}} = \hat{\alpha}_{\text{overall}} + \hat{\beta}_{\text{overall}} Z$ with a model that uses no external information at all, namely the re-estimated model $\hat{Y}_{\text{new}} = \hat{\alpha}_{\text{new}} + \hat{\beta}_{1,\text{new}} X_1 + \cdots$

$+\hat{\beta}_{p,\text{new}}X_p$. The validity of the re-calibrated model with estimates for α_{overall} and β_{overall} can be compared to that of the re-estimated model with new estimates for all regression coefficients using an F test:

$$F_{\text{cal}} = \frac{\sum(\hat{Y}_{\text{new}} - \hat{Y}_{\text{cal}})^2/(p-1)}{\sum(Y - \hat{Y}_{\text{new}})^2/(m-p-1)}$$

where p indicates the number of predictors and m indicates the number of patients in the validation data set. Observe that the difference in dimension between the calibration model and the fully new model is equal to $p-1$. If this test is non-significant, we may assume that the re-calibrated model is reasonable. A significant test result indicates that regression coefficients need to be re-estimated.

2.3. Shrinkage towards re-calibrated coefficients

The generalized shrinkage model can nicely be applied to shrink the re-estimated model towards the re-calibrated model. The heuristic shrinkage factor is given by $\hat{c}_{\text{cal}} = (F_{\text{cal}}-1)^+/F_{\text{cal}}$ with compromise model $Y_{\text{pred}} = (1-\hat{c}_{\text{cal}})\hat{Y}_{\text{cal}} + \hat{c}_{\text{cal}}\hat{Y}_{\text{new}}$. In terms of the regression coefficients, this combination of re-calibration and shrinkage can be written $\hat{\beta}_{\text{shrunk+cal}} = \hat{\beta}_{\text{cal}} + \hat{c}_{\text{cal}}(\hat{\beta}_{\text{new}} - \hat{\beta}_{\text{cal}})$.

2.4. Extension to GLMs

For GLMs such as logistic regression or Poisson regression, the heuristic shrinkage factor as used above is obtained by replacing the F-statistic by the standardized chi-square test-statistic, that is $\tilde{F} = \chi^2_{\text{model}}/\text{d}f$, where $\text{d}f$ is the difference in degrees of freedom between the tested model and the null-model. Shrinkage is slightly more complicated than for simple linear regression. Due to the non-linear effects, the intercept has to be adjusted if a model is shrunken towards the simple null-model. We propose to apply (generalized) shrinkage to the regression coefficients and to re-estimate the constant afterwards in order to have the overall mean of the shrunken model equal to the sample mean in the new data set. Details of the application in our case study are presented in Section 4.

3. CASE STUDY WITH LOGISTIC REGRESSION ANALYSIS

3.1. Patient data

We analysed 30-day mortality in a large data set of patients with acute myocardial infarction (GUSTO-I) [17, 18]. This data set has been used before to study methodological aspects of regression modelling [19–22]. In brief, it consists of 40 830 patients, of whom 2851 (7.0 per cent) had died within 30 days. Patients were randomized between four thrombolytic therapy strategies, but treatment allocation was ignored in the present analysis. Patients were accrued between 1990 and 1993 at one of 1082 participating hospitals in 14 countries.

Within the GUSTO-I data set, different geographical regions were distinguished. Eight in the United States (U.S.) [23]; four in Europe (based on combinations of neighbouring countries); Canada; and Australia/New Zealand. For the present analysis we merged 2 relatively small regions in the U.S. (Mid-South and Mid-West), such that 13 regions were created which

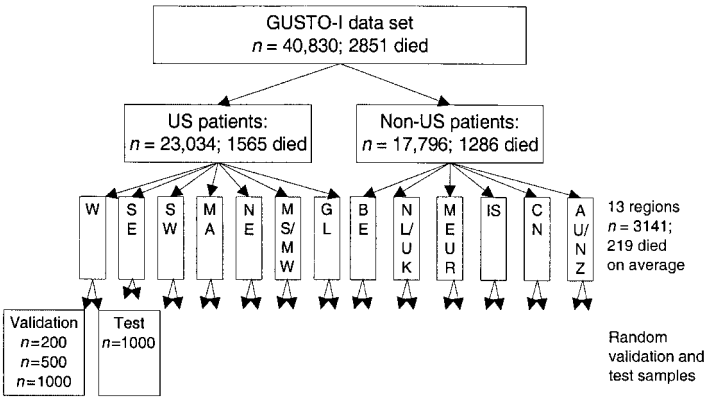

Figure 1. Schematic presentation of the sampling design of the simulation study. The GUSTO-I data was split in 13 regions. The seven U.S. regions were West (W), South-East (SE), South-West (SW), Massachusetts (MA), New England (NE), Mid-South/Mid-West (MS/MW), and Great Lakes (GL). The six non-U.S. regions were Belgium (BE), the Netherlands/United Kingdom (NL/UK), middle Europe—including France, Spain, Germany, Poland—(MEUR), Israel (IS), Canada (CN), and Australia/New Zealand (AU/NZ). Updating methods 1–8 were applied in random samples from each region with sizes of 200, 500, or 1000 patients. Updated models were tested in independent test samples with 1000 patients from the same region as where the validation sample originated from.

included at least 2000 patients each (Figure 1). On average, regions contained 3141 patients and 219 deaths.

3.2. *Prediction models*

We concentrated on a previously published model ('TIMI-II model') which included eight dichotomous predictors: shock, age >65 years, high risk (anterior infarct location of previous MI), diabetes, hypotension (systolic blood pressure <100 mmHg), tachycardia (pulse >80), relief of chest pain >1 h, female gender [24]. The outcome was 42-day mortality, in contrast to 30-day mortality in GUSTO-I. The model was previously developed with backward stepwise selection methods in data from the TIMI-II trial, which included 3339 patients treated in 50 U.S. centres between 1986 and 1988 [25]. In addition to these eight predictors we considered eight previously identified predictors for inclusion in an updated model: height, weight, hypertension, smoking, hypercholesterolaemia, previous angina, family history, and ST elevation in >4 leads [26, 27]. These eight additional predictors were not independent from the eight TIMI-II predictors. Of the 64 correlations considered, most were weak. Absolute Pearson correlation coefficients ($|r|$) were <0.10 for 51 pairs, $0.10<|r|<0.20$ for seven pairs, $0.20<|r|<0.4$ for four pairs, and $|r|\geqslant 0.4$ for two pairs (height and sex, $r=-0.65$; ST elevation in >4 leads and 'high risk', $r=0.39$).

We also considered the situation that a smaller sample size would have been used for construction of the TIMI-II model: $n=500$ instead of $n=3339$. Hereto, a data set was generated with a marginal distribution of predictors as in the TIMI-II population, and predicted probabilities according to the original TIMI-II model. Dichotomous outcomes were generated from a random uniform distribution with the predicted probability as the cut-off. Hence,

the estimated logistic regression coefficients of the eight predictors varied in each generated data set.

For further illustration, two other previously developed prediction models were validated in GUSTO-I. One model was based on a relatively small sample from one coronary care unit in Belgium (477 patients treated between 1977 and 1980). It included 1 dichotomous predictor (anterior infarct location) and 2 continuous predictors (age and left ventricular function) [28]. Another model was based on the relatively large GISSI-2 study ($n = 9720$ patients treated in Italy in the 1980s). It included four predictors with 9 df: age (in five categories), Killip class (in four categories), number of leads with ST elevation (in four categories) and anterior infarct location (dichotomous) [26].

3.3. Simulation studies

Validation samples were randomly drawn from each region in GUSTO-I (Figure 1). These samples provided the data to update the TIMI-II model. Sample sizes were 200, 500 or 1000 patients, with sampling of a constant fraction of survivors and deaths according to the prevalence in a region. The updated model was subsequently tested in sets of 1000 independent patients from the same region as where the validation sample originated from. The procedure was repeated 100 times within each of the 13 regions, resulting in a total of 1300 evaluations for each validation sample size. A similar procedure was followed for smaller development samples: 100 simulations were performed, which included estimation of regression coefficients of the TIMI-II model in development samples of $n = 500$ or 3339, and evaluation in each of the 13 regions with validation samples of $n = 200$, 500, or 1000. This resulted in 1300 evaluations on the test samples of $n = 1000$ for each combination of development and validation sample size.

Updating of the TIMI-II model was further studied in the U.S. patients ($n = 23\,034$) to gain insight in the performance of updating methods with larger validation samples. Validation sample sizes were 1000, 2000, 5000, or 10 000 patients, with testing of the updated models in 10 000 of the remaining U.S. patients. For each validation sample size, 200 evaluation were obtained.

4. UPDATING METHODS

We considered several methods to update a previously defined logistic regression model (Table I). The methods were ordered according to the number of parameters that were estimated for updating of the original model. The first method was not to allow for any updating, that is to keep all regression coefficients fixed at their original value, including the intercept. The linear predictor Z for method 1 (Z_1) was calculated as

$$Z_1 = \alpha_{\mathrm{TIMI}} + \sum_{i \in 1,\ldots,8} \beta_{i,\mathrm{TIMI}} x_i$$

where α_{TIMI} is the intercept and $\beta_{i,\mathrm{TIMI}}$ are the eight regression coefficients that were previously published for the TIMI-II study [25], and x_i the predictor values in samples from the GUSTO-I study [18]. This method provided a reference upon which improvement should be obtained with updating.

Table I. Updating methods considered for the TIMI-II model in the GUSTO-I data.

No.	Updating method	Predictors considered	Parameters tested	Parameters estimated
1	No adjustment	8	0	0
2	Intercept α	8	0	1
3	α + calibration slope β_{overall}	8	0	2
4	$\alpha + \beta_{\text{overall}} + \gamma_{1..8\mid p\leqslant 0.05}$	8	8	2–9
5	$\alpha + \beta_{1..8}$	8	0	9
6	$\alpha + \beta_{\text{overall}} + \gamma_{1..8\mid p\leqslant 0.05} + \beta_{9..16\mid p\leqslant 0.05}$	16	16	2–17
7	$\alpha + \beta_{1..8} + \beta_{9..16\mid p\leqslant 0.05}$	16	8	9–17
8	$\alpha + \beta_{1..16}$	16	0	17

The original model contained eight predictors. Eight additional predictors are considered in methods 6–8. Methods 4 and 6 involve testing of the deviations from recalibrated regression coefficients (γ) with $p\leqslant 0.05$ as the selection criterion.

The second and third methods were simple re-calibration methods. Updating of the intercept intends to correct 'calibration in the large', i.e. to make the average predicted probability equal to the observed overall event rate: $Z_2 = \hat{\alpha} + Z_1$. Hereto we fit a logistic regression model in the validation sample with the intercept α as the only free parameter and the linear predictor based on TIMI-II (Z_1) as an offset variable (i.e. the slope is fixed at unity). In method 3, we update both the intercept α and the overall calibration slope β_{overall} by fitting a logistic regression model in the validation sample with the linear predictor based on TIMI-II as the only covariable: $Z_3 = \hat{\alpha} + \hat{\beta}_{\text{overall}} Z_1$. This method has also been labelled 'logistic calibration' [1].

Methods 4–8 made more structural changes to the model, referred to as 'model revision' [5]. With method 4, we first performed method 3, and then tested whether predictors had an effect that was clearly different in the validation sample. We hereto performed likelihood ratio tests of model extensions in a forward stepwise manner, stopping when $p>0.05$ for each covariable. As a maximum, seven covariables were selected, since β_{overall} was always included in the model. The number of estimated parameters could hence vary between 2 and 9. The linear predictor becomes:

$$Z_4 = \hat{\alpha} + \hat{\beta}_{\text{overall}} Z_1 + \sum_{i\in s} \hat{\gamma}_i x_i$$

where s indicates the selection (maximum 7) out of covariables $1,\ldots,8$, and γ_i the deviation from the re-calibrated coefficient value: $\hat{\gamma}_i = \hat{\beta}_i - \hat{\beta}_{\text{overall}} \beta_{i,\text{TIMI}}$.

With method 5 we fit the TIMI-II model anew:

$$Z_5 = \hat{\alpha} + \sum_{i\in 1,\ldots,8} \hat{\beta}_i x_i$$

where $\hat{\beta}_i$ are the re-estimated coefficients for the eight covariables i as specified in the TIMI-II model. Note that method 4 falls in between method 3 and 5: if selection of $\hat{\beta}_i$ is extremely stringent (p-value of 0), method 4 is equal to method 3 (no individual coefficients re-estimated), and if selection is extremely liberal (p-value of 1), method 4 is equal to method 5 (all individual coefficients re-estimated).

Methods 6–8 consider additional predictors, and might hence be labelled 'model extension' methods. Method 6 is an extension of method 4: we re-calibrate the TIMI-II model with an intercept α and the overall calibration slope β_{overall}, and stepwise test 16 predictors for statistically significant effects. The linear predictor becomes:

$$Z_6 = \hat{\alpha} + \hat{\beta}_{\text{overall}} Z_1 + \sum_{i \in s} \hat{\gamma}_i x_i + \sum_{j \in s2} \hat{\beta}_j x_j$$

where s and $s2$ indicate the selection out of covariables $1, \ldots, 8$ (maximum 7) and additional covariables $9, \ldots, 16$, respectively.

Method 7 is another variant, extending method 5:

$$Z_7 = \hat{\alpha} + \sum_{i \in 1,\ldots,8} \hat{\beta}_i x_i + \sum_{j \in s2} \hat{\beta}_j x_j$$

where $s2$ indicates the selection out of additional covariables $9, \ldots, 16$ that have statistically significant effects in the validation sample.

With method 8 we fit a model with 16 covariables k, i.e. 8 from the TIMI-II model and 8 additional covariables: $Z_8 = \hat{\alpha} + \sum_{k \in 1,\ldots,16} \hat{\beta}_k x_k$.

4.1. Shrinkage variants

We extended methods 4–8 with shrinkage as discussed in Section 2. We calculated the shrinkage factor as

$$\hat{c}_{\text{cal}} = \frac{(\text{Model } \chi^2_{\text{extended-recalibrated}} - \text{d}f)^+}{\text{Model } \chi^2_{\text{extended-recalibrated}}}$$

where the model χ^2 was based on the difference in $-2\log$ likelihood between a model with re-estimated predictors and the recalibrated model, and $\text{d}f$ corresponded to the difference in degrees of freedom of these models. Regression coefficients were shrunken towards their re-calibrated values as obtained with method 3. For the first eight predictors in our study this means that any re-estimated coefficients are shrunken towards the re-calibrated values from TIMI-II ($\beta_{\text{overall}}\beta_{i,\text{TIMI}}$) with methods 4 and 5. The coefficients of the additional eight predictors considered in methods 6–8 were shrunken towards zero since these predictors were not included in the TIMI-II model. The intercept of the shrunken model was re-estimated to ensure that the sum of predicted probabilities equalled the sum of observed outcomes (in our case: deaths). When stepwise regression was applied to select predictors for the model, the degrees of freedom of the candidate predictors was considered in the formula [11, 29].

4.2. Calculations

All calculations were performed with S-plus software (MathSoft, Inc., Seattle, WA, version 2000). We used functions from the Design library for logistic regression (`lrm.fit`) and validation (`val.prob`) [30]. All updating methods were performed within the same samples, smaller validation samples were subsamples of the larger ones, and all updated models were tested on the same independent data. This allows for efficient pairwise comparisons of findings. Dot charts were created to enable a visual comparison of the joint effects of alternative updating methods and sample size [31]. Averages of performance measures were calculated with 5 per cent trimming to improve stability.

5. PERFORMANCE MEASURES

We used a number of performance measure to validate the TIMI-II model and to evaluate updated versions of that model. Validation included an assessment of calibration, discrimination and overall performance.

For calibration, we focused on the intercept and slope framework as originally proposed by Cox [6]. The calibration slope is the estimated regression coefficient β in a logistic regression model with the linear predictor as the only covariate [32]:

$$\text{observed mortality} \sim \alpha + \beta \text{ linear predictor}$$

The observed mortality is coded binary (0/1), and the usual logistic link function is applied. The linear predictor (or prognostic index) was calculated as the linear combination of the regression coefficients from the TIMI-II trial with the values of the covariables for each patient in the test data. Well-calibrated models have an intercept α of zero and a slope β of 1. To quantify miscalibration we used the unreliability index U, which is the difference in $-2\log$ likelihood of a model with both α and β as free parameters and a model with $\alpha = 0$ and $\beta = 1$ [32, 33]. The statistic was scaled by dividing by the number of patients.

For discrimination, we used the concordance statistic c, which is equivalent to the area under the receiver operating characteristic (ROC) curve. C varies between 0.5 and 1.0 for sensible models; the higher the better [1, 32].

As a measure of overall performance, we studied the Brier score or average prediction error. The Brier score is calculated as $\sum(y_i - p_i)^2/n$, where y denotes the observed outcome and p the prediction for subject i in the data set of n subjects [34]. The Brier score is 0 for perfect models. When the predicted mortality is 7.0 per cent for every patient (equal to the average observed mortality in our example), the Brier score is 0.0651.

6. EXAMPLES OF VALIDATION AND UPDATING

6.1. *Validity of the TIMI-II model*

As a first impression of validity we compared the regression coefficients as previously estimated in the TIMI-II trial ($n = 3339$) to those in GUSTO-I ($n = 40\,830$) (Table II). We note that the coefficients were reasonably similar, although the coefficients of age and hypotension were somewhat larger in GUSTO-I, and those of shock, high risk, and especially diabetes smaller.

We further studied the estimated coefficients in smaller parts of the GUSTO-I data set, which illustrates the combined effect of regional and sampling variability. A total of 23 034 patients were included from the U.S. Within the U.S., 2188 patients were treated in 55 hospitals in the Western region of the U.S. Of these, 429 were described in detail in a previously report (noted as 'sample 5') [22]. For diabetes, the (non-significant) coefficient was close to zero in the West region and negative in sample 5. Also, the effect of sex vanished in the smallest sample. We tested for evidence of regional variability among the 13 regions. The intercept depended clearly on region ($p < 0.001$), while the effects of the eight predictors did not (all interaction terms of predictors *region had p-values > 0.05).

Table II. Logistic regression coefficients ± standard error in the TIMI-II data and in parts of the GUSTO-I data.

Variables	TIMI-II $n = 3339$	GUSTO-I Total $n = 40\,830$	GUSTO-I U.S. patients $n = 23\,034$	GUSTO-I W region $n = 2188$	GUSTO-I sample 5 $n = 429$
Shock	1.79 ± 0.29	1.60 ± 0.08	1.56 ± 0.11	2.39 ± 0.41	2.96 ± 0.92
Age > 65	0.99 ± 0.18	1.43 ± 0.05	1.34 ± 0.06	1.64 ± 0.22	1.37 ± 0.49
High risk	0.92 ± 0.26	0.71 ± 0.04	0.70 ± 0.06	0.85 ± 0.21	0.76 ± 0.50
Diabetes	0.74 ± 0.19	0.28 ± 0.05	0.31 ± 0.07	0.07 ± 0.25	-0.11 ± 0.64
Hypotension	0.69 ± 0.27	1.19 ± 0.06	1.19 ± 0.07	1.22 ± 0.25	1.39 ± 0.57
Tachycardia	0.59 ± 0.16	0.62 ± 0.04	0.61 ± 0.06	0.65 ± 0.20	0.88 ± 0.49
Time to relief	0.53 ± 0.20	0.50 ± 0.05	0.51 ± 0.06	0.26 ± 0.21	0.68 ± 0.54
Sex	0.47 ± 0.19	0.43 ± 0.04	0.47 ± 0.06	0.62 ± 0.20	-0.04 ± 0.51
Intercept	-4.47 ± 0.35	-4.82 ± 0.06	-4.84 ± 0.09	-5.09 ± 0.30	-5.19 ± 0.72

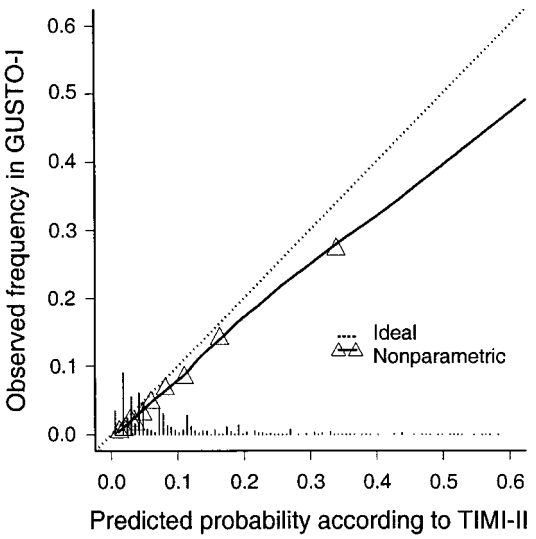

Figure 2. Calibration plot of the TIMI-II model (developed in $n = 3339$) to predict 30-day mortality after acute myocardial infarction in GUSTO-I ($n = 40\,830$). The solid line represents a non-parametric smooth curve for the relation between predicted probability and observed frequency. Perfect calibration is represented by the dotted line through the origin with slope equal to 1. Triangles are based on deciles of patients with similar predicted probabilities. The distribution of predicted probabilities is shown above the x-axis (vertical lines). We note that the predicted risks are systematically too high.

The validity of the TIMI-II model for the GUSTO-I patients is further illustrated in a calibration plot (Figure 2). We note that the observed mortality is systematically lower than predicted. This might at least partly be attributed to the slight difference in outcome definition (30-day mortality in GUSTO-I versus 42-day mortality in TIMI-II).

6.2. *Illustration of updating methods*

We illustrate the application of the updating methods 2–4 in Table III. Corresponding to the observed miscalibration in Figure 2, the intercepts were negative (around -0.3) when method 2 was applied, with somewhat more extreme estimates in smaller validation sets. The corresponding odds ratios were between 0.63 (sample 5, $e^{-0.47}\, p = 0.03$) and 0.76 (total GUSTO-I data set, $e^{-0.28}\, p < 0.001$), indicating that the predicted probabilities were approximately 1.3–1.6 times too high. The calibration slopes were close to 1 (method 3).

Method 4 updates the original model as in method 3 plus estimation of coefficients that were clearly different from overall re-calibrated values. We found that the effects of age, high risk, diabetes, hypotension, and tachycardia were clearly different in the total GUSTO-I data set. No statistically significant deviations were observed in the smallest sample, obviating a clear need for re-estimation of individual coefficients (Table III).

The results of method 5, fitting the model anew, were illustrated before (Table II). For updating methods 6–8, eight additional predictors were considered. In a 16-predictor model, these eight were each statistically significant ($p < 0.01$) in the full GUSTO-I data set ($n = 40\,830$) and the U.S. part ($n = 23\,034$), but their predictive effects were smaller than those of the eight predictors from the TIMI-II model. Additional predictors with statistically significant effects were weight and ST elevation in the West region, and none in the smallest sample.

Table III. Illustration of updating of the TIMI-II model in parts of the GUSTO-I data according to calibration methods (methods 2 and 3) and model revision with statistically significant different coefficients (method 4).

	GUSTO-I total $n = 40\,830$	GUSTO-I U.S. patients $n = 23\,034$	GUSTO-I region 1 $n = 2188$	GUSTO-I sample 5 $n = 429$
Re-calibration: method 2				
α: intercept	-0.28 ± 0.02	-0.34 ± 0.03	-0.36 ± 0.09	-0.47 ± 0.22
Re-calibration: method 3				
α: intercept	-0.28 ± 0.03	-0.39 ± 0.05	-0.10 ± 0.16	-0.26 ± 0.47
β_{overall}: calibration slope	0.99 ± 0.02	0.98 ± 0.03	1.13 ± 0.09	1.11 ± 0.22
Revision: method 4*				
α: intercept	-0.76 ± 0.15	-0.62 ± 0.17	-0.25 ± 0.36	-0.26 ± 0.47
β_{overall}: calibration slope	0.91 ± 0.04	0.94 ± 0.04	1.14 ± 0.12	1.11 ± 0.22
γ_1: shock	$+0$	$+0$	$+0$	$+0$
γ_2: age > 65	$+0.53 \pm 0.06$	$+0.42 \pm 0.07$	$+0.49 \pm 0.24$	$+0$
γ_3: high risk	-0.12 ± 0.06	-0.17 ± 0.07	$+0$	$+0$
γ_4: diabetes	-0.39 ± 0.06	-0.38 ± 0.08	-0.79 ± 0.27	$+0$
γ_5: hypotension	$+0.56 \pm 0.07$	$+0.52 \pm 0.08$	$+0$	$+0$
γ_6: tachycardia	$+0.09 \pm 0.05$	$+0$	$+0$	$+0$
γ_7: time of relief	$+0$	$+0$	$+0$	$+0$
γ_8: sex	$+0$	$+0$	$+0$	$+0$

*The updated regression coefficients β_i can be calculated as $\beta_{\text{overall}}\beta_{i,\text{TIMI}} + \gamma_i$, where $\beta_{i,\text{TIMI}}$ are considered known. Calculation of the corresponding standard errors of these updated coefficients would require an estimate of the covariance between β_{overall} and γ_i.

Table IV. Apparent performance of updated versions of the TIMI-II model in parts of the GUSTO-I data.

	Method	GUSTO-I total $n = 40\,830$	GUSTO-I U.S. patients $n = 23\,034$	GUSTO-I Region 1 $n = 2188$	GUSTO-I sample 5 $n = 429$
Parameters estimated	1	0	0	0	0
	2	1	1	1	1
	3	2	2	2	2
	4	7	6	4	2
	5	9	9	9	9
	6	17	13	5	2
	7	17	17	11	9
	8	17	17	17	17
Miscalibration (U statistic)	1	0.005	0.007	0.007	0.008
	2	0.000	0.000	0.001	0.001
Discrimination (c statistic)	1	0.782	0.780	0.795	0.776
	2	0.782	0.780	0.795	0.776
	3	0.782	0.780	0.795	0.776
	4	0.793	0.791	0.810	0.776
	5	0.793	0.790	0.819	0.793
	6	0.802	0.800	0.819	0.776
	7	0.802	0.800	0.828	0.793
	8	0.802	0.800	0.830	0.851
Overall performance (Brier score)	1	0.059	0.058	0.052	0.047
	2	0.058	0.057	0.051	0.047
	3	0.058	0.057	0.051	0.046
	4	0.058	0.057	0.051	0.046
	5	0.058	0.057	0.051	0.044
	6	0.057	0.056	0.051	0.046
	7	0.057	0.057	0.050	0.044
	8	0.057	0.057	0.050	0.040

Results are shown for methods 1–8, except for miscalibration (U statistic zero for methods 3–8).

6.3. *Apparent performance*

In Table IV we show the apparent performance of the updated models from Tables II and III. When no updating was performed, the miscalibration (as quantified by a scaled chi-square statistic) was 0.0046 in the total GUSTO-I data and slightly larger in the smaller samples (around 0.0007). When the intercept was updated, the unreliability was close to zero. When the model was updated more extensively (methods 3–8), the apparent unreliability was by definition zero.

The c statistic of the TIMI-II model was around 0.78 with methods 1–3. Updating of some (method 4) or all (method 5) of the coefficients led to a somewhat higher apparent discriminative ability (c around 0.80 in the larger samples). The extension of the TIMI-II model with more predictors increased the apparent discriminative ability further, although the increase was small in the total GUSTO-I data set.

The Brier score showed rather small differences between updating methods in the larger samples. A substantial decrease was noted in the smallest sample with more complex updating methods.

It is well known that the apparent performance may be a severely optimistic estimate of performance in new patients. For illustration, we studied the internal validity of three models as identified with method 3, 5, and 8 for the smallest sample ($n = 429$). Models were developed in 200 bootstrap samples and tested in the original sample to estimate the optimism in apparent performance measures [1]. The optimism was smallest for the 2 parameter model (method 3), and largest with the 17 parameter model (method 8), where discrimination was expected to decrease from 0.851 to 0.770, and the Brier score to increase from 0.040 to 0.048. The highest internal validity was found for method 3, with optimism-corrected estimates of c and Brier score of 0.772 and 0.065. This suggests that a model with updating of fewer parameters may perform better in independent data than a more extensively updated model. This issue was further studied with simulation experiments in samples of varying size.

7. SIMULATION RESULTS

7.1. Updating with small validation samples

We first evaluated updating methods in validation samples with 200, 500, or 1000 patients from each of the 13 geographical regions considered. The average number of deaths was 14, 35 or 70, respectively. Since not all methods were considered realistic approaches in the smallest samples, we did not evaluate methods 6–8 for $n = 200$. The average number of parameters estimated with methods 4, 6, and 7 increased from 2.7, 3.4, and 9.7 with $n = 500$ to 3.1, 5.0, and 10.1 with $n = 1000$, respectively, reflecting the increase in power for selection of statistically significant effects.

The average results of the updating methods are shown in Figure 3 for the eight methods and sample sizes considered, both for unshrunken and shrunken models. The performance of the models was determined on test parts within each of the 13 regions ($n = 1000$, Figure 1). The calibration slope was close to one for methods 1 and 2, corresponding to the results shown in Table III. As expected, the slope was also close to one when it was re-estimated in the validation sample (method 3). More extensive updating led to a slope clearly below one, especially when the full 8 or 16 predictor models were re-estimated in the smaller samples (method 5 with $n = 200$: slope $= 0.59$, method 8 with $n = 500$: slope $= 0.82$). These slopes reflect the need for shrinkage in these situations. The unreliability statistic was lowest when only the intercept was updated. No improvement was obtained by updating of the slope, while the more extensive updating methods led to a further deterioration of reliability. Better results were obtained with the shrinkage methods: the calibration slopes became closer to one, with clear improvements for methods 5, 7 and 8.

Discriminative ability of the TIMI-II model was by definition not affected by the recalibration method 2 or 3, with a c statistic of 0.785. Discrimination was not improved by any of the other updating methods, with a decrease to 0.742 by method 5 in the smallest samples ($n = 200$). For the larger samples ($n = 1000$), shrinkage led to a slight improvement over simple re-calibration, e.g. to 0.787 for methods 5, 7 and 8 with shrinkage.

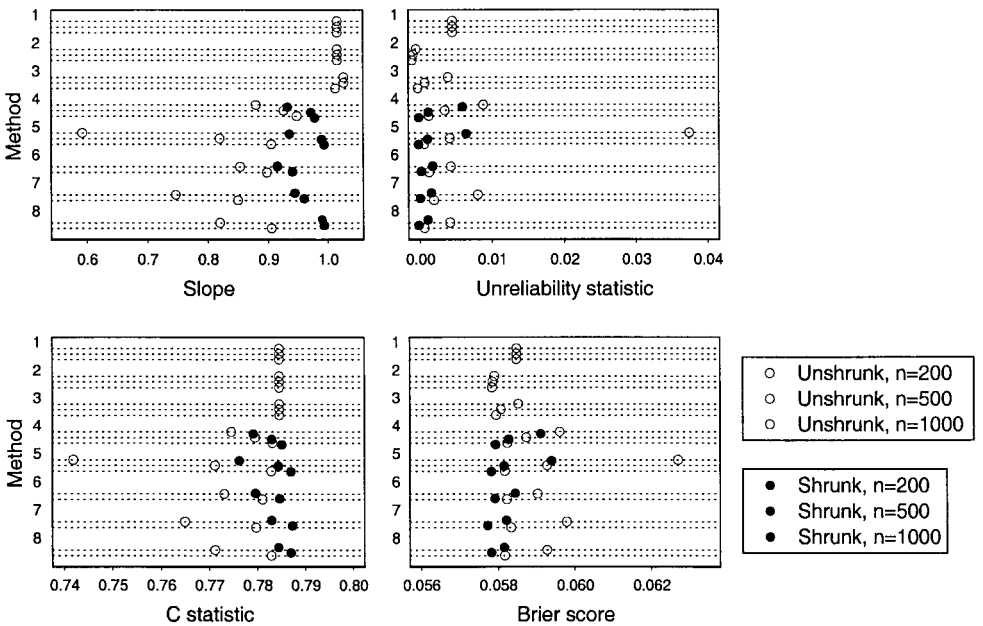

Figure 3. Dotcharts showing the average results for the eight updating methods (numbers 1–8) with or without application of shrinkage in the updating of regression coefficients. For methods 1–5, sample sizes were 200, 500, or 1000 (3 rows). For methods 6–8, sample sizes were 500 or 1000 (2 rows). Validation samples were drawn from 13 regions within the GUSTO-I study. Performance was determined in independent test samples with $n = 1000$, as shown in Figure 1.

The Brier score reflected the patterns observed with calibration and discrimination. Recalibration of the intercept (method 2) led to a low Brier score (good overall performance) for all sample sizes ($n = 200$–1000). With methods 3–8, model performance deteriorated, unless shrinkage was applied and a larger sample size was available for updating (e.g. methods 5, 7, 8 with $n = 1000$).

7.2. Updating with large validation samples

We further evaluated the updating methods with larger validation samples ($n = 1000$–$10\,000$) among the U.S. patients in the GUSTO-I data set (Figure 4). With these larger sample sizes we would expect that especially methods 6–8 might perform better than before. Indeed, method 7 outperformed all other methods with respect to discrimination and overall performance for $n \geqslant 2000$, although the differences were small. Methods 4–8 required shrinkage to obtain a slope of the linear predictor close to one, especially for $n = 1000$–2000.

7.3. Updating of smaller development samples

We also considered updating methods with the TIMI-II model was developed in smaller samples ($n = 500$ instead of $n = 3339$) and validated in samples ranging from $n = 200$ to 1000. With $n = 3339$ for model development, the results were very similar to those of using the original TIMI-II model (results not shown, closely resembling Figure 3). With $n = 500$ for

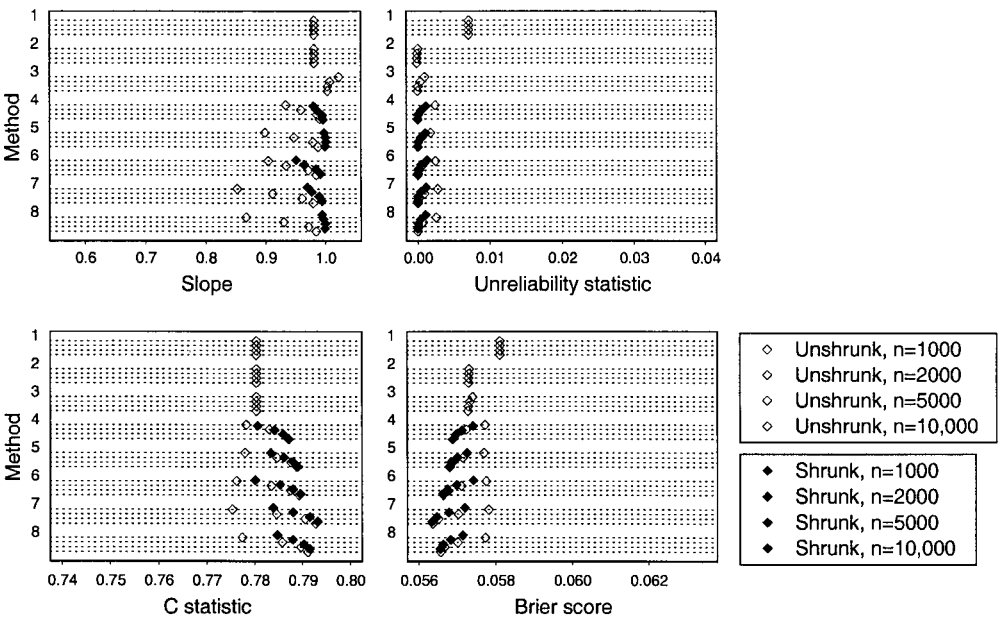

Figure 4. Dotcharts showing the results of simulation studies in the U.S. patients from the GUSTO-I study. Average results are shown for the eight updating methods (numbers 1–8), with or without application of shrinkage in the updating of regression coefficients. Validation sample sizes were 1000, 2000, 5000, or 10 000 (4 rows for each method), with test sample sizes of $n = 10\,000$.

model development, we found that the slope of the linear predictor was around 0.8 (method 1 or 2, Figure 5). This reflects a need for shrinkage, which is consistent with the small development sample size. Updating of the slope (method 3) solved this problem.

The unreliability was considerable for method 1, and when methods 4 and 5 were applied in the smallest validation samples ($n = 200$). Method 3 performed better than 2, which was in contrast to the original TIMI-II model (developed in $n = 3339$).

The discriminative ability was hampered by the smaller size of the development data set (c around 0.75 for methods 1–3, in contrast to around 0.78 for the original TIMI-II model). A more satisfactory performance was obtained with methods 5, 7, or 8 for validation samples sizes $\geqslant 500$, especially when combined with shrinkage.

The Brier score was reasonable for method 3, but could further be reduced with other methods when larger validation samples were available ($n \geqslant 500$). Shrinkage methods led to better results for all performance measures, especially for small validation samples.

7.4. Other prediction models

Simple re-calibration (method 3) was also considered for two other prediction models: the Belgium model and the GISSI-2 model. With validation in GUSTO-I we found calibration slopes reasonably close to 1 for both models (1.26 and 0.87, Figure 6). These slopes are not easy to grasp from the graph with probability scales, but are evident when log odds scales are used. The intercepts were negative when the linear predictor was fixed at unity (Method

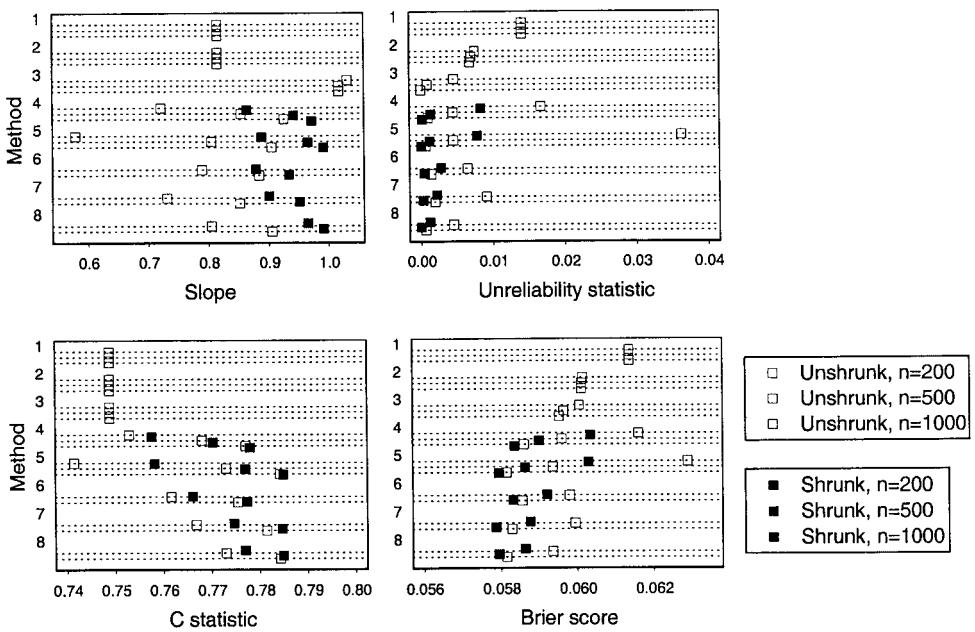

Figure 5. Dotcharts showing the results of simulation studies with smaller development samples ($n = 500$ instead of $n = 3339$ for the original TIMI-II model as shown in Figure 3). Average results are shown for the eight updating methods (numbers 1–8), with or without application of shrinkage in the updating of regression coefficients. For methods 1–5, validation samples contained 200, 500, or 1000 patients (3 rows). For methods 6–8, sample sizes were 500 or 1000 (2 rows). Performance was determined in independent test samples with $n = 1000$, as shown in Figure 1.

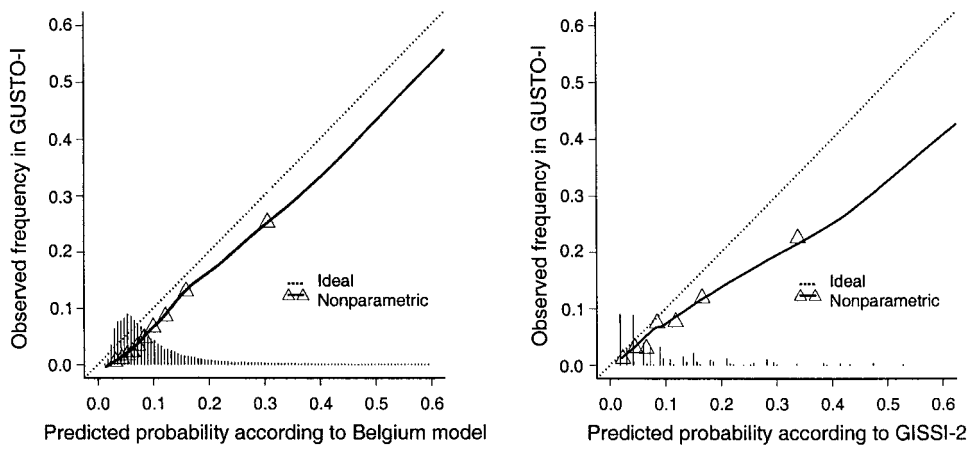

Figure 6. Calibration plots of the Belgium model (developed in $n = 477$) and the GISSI-2 model (developed in $n = 9720$) to predict 30-day mortality in GUSTO-I ($n = 40\,830$).

2, $\alpha - 0.46$ and -0.44, respectively), reflecting a lower mortality in GUSTO-I compared to the earlier series.

8. DISCUSSION

The updating methods considered in this study differ in their extensiveness of adjusting a previously published model. Methods that focus on calibration require estimation of only one or two parameters (intercept, or intercept and overall slope) and are hence rather robust. However, they rely on the assumption that the relative strength of predictors was approximately similar in the development and validation samples. This assumption was reasonable in our case study where a model developed in the TIMI-II study was validated in the GUSTO-I study. Therefore improvements upon simple re-calibration methods were not possible with small validation data sets. Further model revision was performed by re-estimating the regression coefficients for all predictors or only for those coefficients that were clearly different in the validation sample. Such revision was beneficial if the development sample was small and if the validation sample was of similar size or larger than the development sample. A more extensive revision is to consider additional covariables. This made limited sense in our case study, where the initial set of predictors already contained strongly predictive covariables, which were modestly correlated with the additional predictors. Furthermore, the more extensive revisions could harm the quality of the updated model for future patients in the validation setting. The potential for harm by a priori 'reasonable' updating strategies is an important finding of our study.

For all methods we assume that it is considered reasonable from a clinical point of view to apply the previously published model in a new patient sample. The investigated updating methods however decreased in the assumptions made about the similarity between development and validation populations. Method 1 ('no updating') assumes that nothing has changed between the populations and that the previously estimated regression coefficients (including the intercept) are correct for the validation population. Method 2 allows for a difference in severity of the patients ('case mix'), not reflected in the model parameters. We may think of a missed predictor in the model, for example related to the referral pattern of patients, or a difference in treatment that affects all patients in a similar way. Method 3 ('re-calibration') allows for a generally smaller or larger effect of the predictors. It assumes that the relative effects of the predictors are similar however. Methods 4 and 5 relax the latter assumption by allowing adjustments of the relative effects. Effects of predictors might e.g. change when a treatment specifically affects patients with certain characteristics, or when definitions of predictors differ. Re-estimation of individual coefficients may also be necessary when the correlations between predictors are different in the validation setting compared to the development setting. Methods 6–8 allow for more predictors to be included, which is related to the general issue of prognostic model development. A key problem of such more extensive model revisions is that overoptimistic predictions may be constructed, especially when a small validation data set was used for updating. This overoptimism could be reduced by applying heuristic shrinkage factors, in the same spirit as previous proposals [10, 11]. Shrinkage towards the re-calibrated coefficient values worked well for pre-specified models (methods 5 and 8), but also for methods that involved testing for a different effect of predictors (methods 4, 6, and 7). Shrinkage not only led to an improvement in calibration, but also in discrimination.

We may view model revision with shrinkage as a way to provide regression coefficients that are updated to the extent supported by the data. This is noted from the heuristic formula to calculate the shrinkage factor, where the fit of a re-estimated and re-calibrated model is compared. Recently, methods have been proposed that shrink coefficients to zero, thus leading to selection of parameters through shrinkage (Garotte [12], Lasso [13]). These approaches might work even better for methods 4, 6, and 7, but are computationally more burdensome to apply.

Our case study supported re-calibration for validation data sets that were smaller than the development data set. Also, re-calibration was the main issue in a number of other prediction problems, rather than revision of the full model. Especially, the intercept may be different in other settings, while the slope remains close to unity. Examples include a survival model for kidney graft survival, which had a slope of the linear predictor close to one in more recently transplanted patients (0.97) [4]. In this study, between centre heterogeneity was modelled with a mixed effect approach, assuming identical predictor effects across the centres. Further, the slope was 0.985 when a externally developed risk-adjustment model was applied in a data set of patients undergoing coronary artery bypass grafting [7]. Finally, in validation studies of a prediction rule for testicular cancer patients, we found slopes of 0.97 and 0.91 [35–37].

On the other hand, examples have been given of models which calibrated and discriminated poorly when validated externally [3]. These models were typically constructed in small data sets, where a large set of candidate predictors was considered. Standard procedures were followed that lead to suboptimal predictive models, e.g. stepwise selection [16, 21, 33]. Shrinkage was not applied, and internal validity was not assessed adequately. The resulting models had calibration slopes that were far below 1 in external data [3], in line with our results with smaller development data sets. In addition to adjustment of the intercept, re-estimation of the slope is required, although one may doubt whether such models could be salvaged that simply. Note that a slope correction could partly be seen as '*post hoc* shrinkage', when some sort of shrinkage was not performed earlier for a model constructed in a small data set.

We further note that a substantial size will be required for a validation sample to quantify (in)validity in a reliable way, i.e. with enough power to exclude important miscalibration or a substantial decrease in discriminative ability. In this respect, it is unfortunate that guidelines for sample sizes in validation studies of regression models are lacking thus far. If a large validation data set is available, a poorly performing model can actually largely be discarded. It might then be most sensible to define a new model, with predictors selected based on subject knowledge (other studies, expert opinion), and with estimation of regression coefficients including shrinkage [1, 16, 21].

A limitation of the GUSTO-I example is that the slope of the previously developed TIMI-II model was very close to 1. This implies that limited improvement was possible with respect to overall calibration. On the other hand, individual predictors had significantly different effects in GUSTO-I compared to TIMI-II. A better incorporation of these effects should be reflected in performance measures such as the Brier score, but this was only noticeable with the larger validation samples (Section 6). Further, two other previously developed models performed rather similar to the TIMI-II model with respect to calibration slope (Figure 6).

In conclusion, simple re-calibration may be a reasonable strategy to obtain a valid model for another population, especially when a relatively small validation data set is available for updating. Updating of all coefficients may then harm predictive performance. We propose to

perform shrinkage of updated coefficients towards their re-calibrated values when a relatively large validation data set is available.

ACKNOWLEDGEMENTS

We would like to thank Kerry L. Lee, Duke Clinical Research Institute, Duke University Medical Center, Durham NC, and the GUSTO investigators for making the GUSTO-I data available for analysis. The research of Dr Steyerberg has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

REFERENCES

1. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.
2. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999; **130**:515–524.
3. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; **19**:453–473.
4. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine* 1995; **14**:1999–2008.
5. Van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 2000; **19**:3401–3415.
6. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45**:562–565.
7. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* 1997; **16**:2645–2664.
8. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or Remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation* 1999; **99**:2098–2104.
9. Miller Me, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Statistics in Medicine* 1991; **10**:1213–1226.
10. Copas JB. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society Series B* 1983; **45**:311–354.
11. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Statistics in Medicine* 1990; **9**:1303–1325.
12. Breiman L. Better subset regression using the nonnegative Garotte. *Technometrics* 1995; **37**:373–384.
13. Tibshirani R. Regression and shrinkage via the Lasso. *Journal of the Royal Statistical Society Series B* 1996; **58**:267–288.
14. Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of shrinkage techniques in logistic regression analysis: a case study. *Statistica Neerlandica* 2001; **55**:76–88.
15. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Statistics in Medicine* 1993; **12**:717–736.
16. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 1986; **5**:421–433.
17. GUSTO-I Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *New England Journal of Medicine* 1993; **329**:673–682.
18. Lee KL, Woodlief LH, Topol EJ, Weaver WD, Betriu A, Col J, Simoons M, Aylward P, Van de Werf F, Califf RM. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. *Circulation* 1995; **91**:1659–1668.
19. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. *Statistics in Medicine* 1998; **17**:2501–2508.
20. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology* 1999; **52**:935–942.
21. Steyerberg EW, Eijkemans MJ, Harrell Jr FE, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine* 2000; **19**:1059–1079.
22. Steyerberg EW, Eijkemans MJ, Van Houwelingen JC, Lee KL, Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine* 2000; **19**:141–160.

23. Pilote L, Califf RM, Sapp S, Miller DP, Mark DB, Weaver WD, Gore JM, Armstrong PW, Ohman EM, Topol EJ. Regional variation across the United States in the management of acute myocardial infarction. GUSTO-1 investigators. Global utilization of streptokinase and tissue plasminogen activator for occluded coronary arteries. *New England Journal of Medicine* 1995; **333**:565–572.
24. Mueller HS, Cohen LS, Braunwald E, Forman S, Feit F, Ross A, Schweiger M, Cabin H, Davison R, Miller D *et al.* Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction. Analyses of patient subgroups in the thrombolysis in myocardial infarction (TIMI) trial, phase II. *Circulation* 1992; **85**:1254–1264.
25. TIMI-II study group. Comparison of invasive and conservative strategies after treatment with intravenous tissue plasminogen activator in acute myocardial infarction. Results of the thrombolysis in myocardial infarction (TIMI) phase II trial. *New England Journal of Medicine* 1989; **320**:618–627.
26. Maggioni AP. Maseri A, Fresco C, Franzosi MG, Mauri F, Santoro E, Tognoni G. Age-related increase in mortality among patients with first myocardial infarctions treated with thrombolysis. The Investigators of the Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico (GISSI-2). *New England Journal of Medicine* 1993; **329**:1442–1448.
27. Maynard C, Weaver WD, Litwin PE, Martin JS, Kudenchuk PJ, Dewhurst TA, Eisenberg MS, Hallstrom AP, Chambers J. Hospital mortality in acute myocardial infarction in the era of reperfusion therapy (the Myocardial Infarction Triage and Intervention Project). *American Journal of Cardiology* 1993; **72**:877–882.
28. Dubois C, Pierard LA, Albert A, Smeets JP, Demoulin JC, Boland J, Kulbertus HE. Short-term risk stratification at admission based on simple clinical data in acute myocardial infarction. *American Journal of Cardiology* 1988; **61**:216–219.
29. Harrell FE. *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer: New York, 2001.
30. Harrell FE. Design library. 2000; http://lib.stat.cmu.edu/S/Harrell/Design.html.
31. Cleveland WS. *The Elements of Graphing Data*. Wadsworth Advanced Books and Software: Monterey, 1985.
32. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CL. Validation of probabilistic predictions. *Medical Decision Making* 1993; **13**:49–58.
33. Harrell Jr FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 1984; **3**:143–152.
34. Arkes HR, Dawson NV, Speroff T, Harrell Jr FE, Alzola C, Phillips R, Desbiens N, Oye RK, Knaus W, Connors Jr AF. The covariance decomposition of the probability score and its use in evaluating prognostic estimates. *Medical Decision Making* 1995; **15**:120–131.
35. Steyerberg EW, Vergouwe Y, Keizer HJ, Habbema JDF. Residual mass histology in testicular cancer: development and validation of a clinical prediction rule. *Statistics in Medicine* 2001; **20**:3847–3859.
36. Steyerberg EW, Gerl A, Fossa SD, Sleijfer DT, de Wit R, Kirkels WJ, Schmeller N, Clemm C, Habbema JD, Keizer HJ. Validity of predictions of residual retroperitoneal mass histology in nonseminomatous testicular cancer. *Journal of Clinical Oncology* 1998; **16**:269–274.
37. Vergouwe Y, Steyerberg EW, Foster RS, Habbema JD, Donohue JP. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. *Journal of Urology* 2001; **165**:84–88.

