Tutorial in Biostatistics

Statistics in Medicine

Received 21 November 2012, Accepted 17 July 2013 Published online 23 August 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5941

Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers

Peter C. Austin[a,b,c,d*†] **and Ewout W. Steyerberg**[e]

Predicting the probability of the occurrence of a binary outcome or condition is important in biomedical research. While assessing discrimination is an essential issue in developing and validating binary prediction models, less attention has been paid to methods for assessing model calibration. Calibration refers to the degree of agreement between observed and predicted probabilities and is often assessed by testing for lack-of-fit. The objective of our study was to examine the ability of graphical methods to assess the calibration of logistic regression models. We examined lack of internal calibration, which was related to misspecification of the logistic regression model, and external calibration, which was related to an overfit model or to shrinkage of the linear predictor. We conducted an extensive set of Monte Carlo simulations with a locally weighted least squares regression smoother (i.e., the loess algorithm) to examine the ability of graphical methods to assess model calibration. We found that loess-based methods were able to provide evidence of moderate departures from linearity and indicate omission of a moderately strong interaction. Misspecification of the link function was harder to detect. Visual patterns were clearer with higher sample sizes, higher incidence of the outcome, or higher discrimination. Loess-based methods were also able to identify the lack of calibration in external validation samples when an overfit regression model had been used. In conclusion, loess-based smoothing methods are adequate tools to graphically assess calibration and merit wider application. © 2013 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

Keywords: logistic regression; prediction; calibration; graphical methods; prediction models

1. Introduction

Predicting the occurrence of a binary outcome is important in medicine and clinical epidemiology, in health services research, and in population health research. Common binary outcomes include perioperative death versus perioperative survival, presence of disease versus absence of disease, and occurrence versus absence of complications following surgery. Applied health researchers are frequently interested in developing models or algorithms to predict the occurrence of a binary outcome. While logistic regression is the most commonly-used statistical method in the biomedical literature for predicting the probability of the occurrence of a binary outcome, other methods that can be used include regression trees or tree-based ensemble methods such as bagged regression trees, random forests, or boosted regression trees [1–5].

An important component to the development and validation of predictive models is the assessment of their predictive accuracy. There are two primary aspects to this assessment: assessment of discrimination and assessment of calibration [6–8]. Discrimination refers to how well the model discriminates

[a]*Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*
[b]*Institute of Health Management, Policy and Evaluation, University of Toronto, Toronto, Ontario, Canada*
[c]*Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada*
[d]*Sunnybrook Research Institute, Sunnybrook Health Sciences Centre, Toronto, Canada*
[e]*Department of Public Health, Erasmus MC – University Medical Center Rotterdam, The Netherlands*
**Correspondence to: Peter C. Austin, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5, Canada.*
[†]*E-mail: peter.austin@ices.on.ca*

between those who did and did not experience the event or outcome of interest. Discrimination is often evaluated using the c-statistic (equivalent to the area under the receiver operating characteristic curve – often abbreviated as AUC).

Calibration refers to the degree of agreement between observed and predictive probabilities, and is often assessed by testing for lack-of-fit [9, 10]. It is important to distinguish between two different types of calibration: internal and external calibration. Internal calibration refers to agreement between observed and predicted probabilities in the sample in which the model was developed. Lack of internal calibration is related to issues of lack of model fit and misspecification of the fitted logistic regression model. External calibration refers to agreement between observed and predicted probabilities in samples external to the sample in which the model was developed. Lack of external calibration can be related to the original model being overfit in the sample in which it was derived, which occurs especially with relatively small sample sizes [7].

We conducted an extensive set of Monte Carlo simulations to assess the performance of a locally weighted least squares regression smoother (i.e., the loess algorithm) for assessing both internal and external calibration. For the assessment of internal calibration, we examined the ability to graphically detect model misspecification in the following scenarios: (i) the omission of a nonlinear term from the fitted model; (ii) the omission of an interaction between a binary covariate and a continuous predictor variable; (iii) misspecification of the link function by using the logistic link function. As secondary objectives, we examined the influence of the following factors on graphical assessments of calibration: (i) the c-statistic of the underlying logistic regression model; (ii) the incidence of the binary outcome; and (iii) the sample size. For assessment of external calibration, we examined the ability of graphical methods to detect the lack of calibration due to use of a regression model that had been overfit in the derivation sample. Finally, we compared the performance of loess-based methods of assessing calibration with a method based on comparing observed versus predicted probabilities across the ten deciles of risk [9, 10].

2. Calibration

2.1. *Statistical tests for calibration or goodness-of-fit*

Several statistical tests have been proposed to assess model calibration or goodness-of-fit. Hosmer and Lemeshow proposed a goodness-of-fit test on the basis of dividing the sample into strata on the basis of the predicted probability of the outcome [9, 10]. In practice, subjects are often divided into ten, approximately equally-sized, groups on the basis of the deciles of risk. A chi-squared test is then used to compare the observed versus predicted probability of the outcome across the strata. While the Hosmer–Lemeshow test is based on grouping subjects on the basis of the predicted probability of the outcome, Tsiatis proposed a test on the basis of grouping subjects on the basis of the predictor variables [11]. Le Cessie and van Houwelingen proposed tests of goodness-of-fit on the basis of smoothed residuals [12], while Royston proposed tests to detect nonlinearity that used partial sums of residuals [13]. Stukel proposed a generalized logistic model that permits testing of the adequacy of a fitted logistic regression model [14].

While the tests described previously allow one to formally test the goodness-of-fit of the fitted logistic regression model, other authors have proposed methods to qualitatively assess model calibration. Cox proposed a recalibration framework, in which the observed outcome is compared with the linear predictor [15]. An intercept and slope are then estimated, which are related to calibration-in-the-large and the calibration slope [8]. The former compares the mean predicted probability of the occurrence of the outcome with the mean outcome, while the latter, when used for internal validation, reflects the amount of shrinkage that is necessary to make the model well calibrated for predicting outside the derivation sample [8]. A two-degree of freedom test can be performed to test for miscalibration [16]. Harrell and Lee extended Cox's recalibration framework to allow one to derive indices denoting the lack of calibration, discrimination, and overall quality of prediction [17, 18]. Similarly, Dalton recently extended Cox's recalibration framework to provide for a flexible recalibration of binary prediction models [19]. Furthermore, the use of this method permits the derivation of a relative measure of miscalibration for comparing two competing prediction models.

2.2. *Graphical assessment of calibration*

Copas proposed that regression smoothing methods be used to produce calibration plots in which the relationship between observed and predicted probabilities of the outcome is described graphically [20, 21]. This method has since been advocated by different sets of authors [6, 8, 22]. To implement this

approach, the occurrence of the binary outcome is regressed on the predicted probability of the outcome obtained from the fitted logistic regression model. A common approach is to use a locally weighted scatter plot smoother, such as a locally weighted least squares regression smoother (i.e., the loess algorithm) [8, 18, 23]. Plotting the smoothed regression line allows one to examine calibration across the range of predicted values and to determine if there are segments of the range in which the model is poorly calibrated. Harrell *et al*. complemented this calibration plot with a graphical comparison of the observed versus predicted probabilities of the occurrence of the outcome across different strata of risk [6], while 95% confidence intervals can also be added [7]. Deviations of points from a diagonal line with unit slope indicate lack of calibration. When the distribution of predicted probabilities is also given, stratified by outcome, the plot also visualizes discrimination aspects [6–8].

Hosmer *et al*. compared the statistical power of different test-based methods for assessing internal calibration and goodness-of-fit of logistic regression models [24]. Many of the tests had poor statistical power to detect the omission of a binary covariate and its interaction with a continuous predictor variable. Similarly, power was often suboptimal to detect the omission of a quadratic term from a logistic regression model. While much attention has been focused on the performance of different statistical tests for assessing model calibration, there is a paucity of research on the performance of graphical methods for assessing calibration.

3. Design of Monte Carlo simulations

We used an extensive set of Monte Carlo simulations to examine the ability of graphical methods based on locally weighted least squares to assess internal and external calibration of logistic regression models. Three distinct sets of simulations were conducted. First, when the model was correctly specified, we examined the effect of the c-statistic of the logistic regression model and the effect of the incidence of the outcome on graphical assessment of internal calibration. Second, we examined the examined the ability of graphical methods to detect the lack of internal calibration when the regression model had been misspecified. Third, we examined the ability to detect the lack of external calibration due to estimation of an overfit regression model.

Regression smoothing was carried out using the loess function in the R statistical programming language [25]. A key parameter is the span parameter, which denotes the width of the window around each subject such that all subjects within that window are used to fit the weighted least squares regression line used to obtain a prediction for a given subject [23]. In the appendix available as supplemental online material, we have provided a brief comparison of the loess function and the lowess function in R. We also provide a brief examination of the effect of the choice of the span parameter on the assessment of internal calibration. We found that the default value of the span parameter in R (0.75) performed well for assessing calibration and this value was used in all subsequent simulations.

3.1. Effect of c-statistic and incidence of the outcome on the assessment of calibration

The first set of Monte Carlo simulations was designed to examine the effect of the c-statistic of the logistic regression model, and the incidence of the outcome on the graphical assessment of internal calibration. We assumed that the fitted logistic regression model had been correctly specified.

3.1.1. Effect of c-statistic. For each subject in the simulated dataset, a binary outcome was simulated from a Bernoulli distribution with subject-specific parameter p_i, where $\text{logit}(p_i) = \alpha_0 + \alpha_1 x_i$, and $x_i \sim N(0, 1)$. We simulated 50 datasets, each consisting of N subjects. Within each of the simulated datasets we fit a correctly specified logistic regression. For each subject, we determined the predicted probability of the occurrence of the binary outcome by using the fitted logistic regression model. The loess function in R was then used to regress the observed binary outcome on the predicted probability of the outcome. Each of the fitted loess models were then plotted to examine the relationship between the probability of the occurrence of the outcome and the predicted probability of the outcome.

The values of α_0 and α_1 in the data-generating process described previously were chosen so that the outcome would occur for approximately 10% of the subjects and so that logistic regression model would have approximately the desired c-statistic [26]. We used a full factorial design in which we allowed the following two factors to vary. We considered three values of the c-statistic: 0.70, 0.80, and 0.90. We considered three values of the sample size: 500, 1000, and 10,000.

3.1.2. Effect of the incidence of the outcome. We used simulations similar to those described previously to examine the effect of the incidence of the occurrence of the outcome on graphical assessment of calibration. The values of α_0 and α_1 in the data-generating process described previously were chosen so that the outcome would occur for approximately the desired proportion of subjects and so that the c-statistic of the logistic regression model would be approximately 0.80 [26]. We used a full factorial design in which we allowed the following two factors to vary. We considered three values for the percentage of subjects who would experience the event: 1%, 10%, and 50%. We considered three values of the sample size: 500, 1000, and 10,000.

3.2. Assessing internal calibration

In the second set of simulations, we examined the performance of loess-based methods to graphically detect the lack of internal calibration. In particular, we examined the ability of these methods to detect model misspecification in the derivation sample. To do so, we examined scenarios and data-generating processes identical to those examined by Hosmer *et al*. in a study that compared different goodness-of-fit tests for logistic regression models [24]. The following types of model misspecification were examined: (i) omission of a quadratic term from the fitted logistic regression; (ii) omission of an interaction term from the fitted logistic regression; and (iii) the misspecification of the link function of the logistic regression model.

3.2.1. Misspecified regression model: omission of a quadratic term. For each subject in the simulated dataset, a binary outcome was simulated from a Bernoulli distribution with subject-specific parameter p_i, where $\text{logit}(p_i) = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2$, and $x_i \sim U(-3, 3)$. As in Hosmer *et al*., we considered five different scenarios, defined by different values of the three regression coefficients. The lack of linearity in the log-odds function increased progressively across the five scenarios. The degree of nonlinearity in each scenario is described in the top-left panel in Figure 3. Using this data-generating process, we simulated 500 datasets, each consisting of N subjects.

In each of the simulated datasets we fit a misspecified logistic regression model: we regressed the occurrence of the binary outcome on a linear term for the continuous predictor variable and omitted the quadratic term. We used the loess function in R to regress the occurrence of the binary outcome on the predicted probability of the occurrence of the binary outcome derived from the misspecified logistic regression model. We then averaged the estimated loess models across the 500 simulated datasets, as well as determining the 2.5th and 97.5th percentiles of the expected probability of the outcome across the range of predicted values. We conducted these simulations in three different settings defined by the sample size: 500, 1000, and 10000.

We compared the performance of the loess-based method for assessing calibration with that of another commonly-used graphical method for assessing calibration. In each of the 500 simulated datasets, we divided subjects into ten, approximately equally-sized groups, according to the deciles of the predicted probability of the occurrence of the outcome as derived from the fitted logistic regression model. Within each of the ten strata of risk, we determined the mean predicted probability of the occurrence of the outcome and the observed probability of the occurrence of outcome.

3.2.2. Misspecified regression model: omission of an interaction. By using methods identical to those described by Hosmer *et al*., we simulated datasets in which the log-odds of the occurrence of the outcome was related to a continuous variable, a binary variable, and an interaction between these two variables. As in their study, $x \sim U(-3, 3)$, and $d \sim Be(0.5)$ (with the binary covariate d being independent of the continuous covariate x). The log-odds of the occurrence of the outcome was determined from the following logistic model: $\text{logit}(p_i) = \alpha_0 + \alpha_1 x_i + \alpha_2 d + \alpha_3 x d$. We considered four scenarios [24], which allowed for progressively more interaction. The magnitude of the interaction in each scenario is described graphically in the top-left panel of Figures 5 and 6. For each of the four scenarios, we simulated 500 datasets, each consisting of N subjects. In each simulated dataset, we fit two misspecified logistic regression models. First, we regressed the occurrence of the binary outcome on only the continuous predictor variable (i.e., both the binary covariate and its interaction with the continuous covariate were omitted – this was the case examined by Hosmer *et al*. [24]). Second, we regressed the occurrence of the binary outcome on both the continuous and binary predictor variables but omitted the interaction

between the two (this case was not examined by Hosmer *et al*., but we examined it for the sake of completeness). The simulations and analyses then proceeded as described previously. We conducted these simulations in three different settings defined by the sample size (N): 500, 1000, and 10000.

3.2.3. Misspecified regression model: misspecification of the link function. We used methods identical to those described by Hosmer *et al*. [24], which in turn were based on Stukel's generalized logistic model [14], to examine the ability to detect misspecification of the link function. We simulated a single continuous predictor variable: $x \sim U(-3, 3)$, and used the function $\eta(x) = 0.8x$ as the linear predictor variable. We considered five scenarios. The first two were scenarios in which the generalized logistic function had almost the same shape as the probit model and the complimentary log-log model. The remaining three settings were selected so that the generalized logistic function had both tails longer than the logistic model, both tails shorter than the logistic model, or an asymmetric model with one tail longer and one tail shorter than the logistic model. The simulations proceeded as described in the previous subsections.

3.3. Assessing external calibration

We conducted two different sets of Monte Carlo simulations to examine the ability of graphical methods to detect the lack of external calibration. The first examined settings in which the true linear predictor in the validation or external sample was shrunken compared with the linear predictor estimated using the internal or derivation sample. The second examined settings in which an overspecified regression model was developed in the derivation sample and then applied to the external validation sample. Both simulations used data-generating processes that were based on the Enhanced Feedback For Effective Cardiac Treatment-Heart Failure (EFFECT-HF) mortality risk model [27]. This model uses 11 variables to predict the risk of death within 1 year of hospitalization for patients admitted to hospital with a diagnosis of heart failure: age, heart rate, respiratory rate, urea level, low sodium, low hemoglobin, history of stroke or transient ischemic attack, dementia, chronic obstructive pulmonary disease, cirrhosis, and cancer.

3.3.1. Shrunken linear predictor. We used data on 8634 patients hospitalized with heart failure from the EFFECT study and regressed the occurrence of death within 1 year of hospitalization on the 11 variables listed previously by using a logistic regression model [28]. For each patient we estimated the linear predictor from the fitted logistic regression model. The mean and variance of the linear predictor in the EFFECT sample were -0.84 and 0.88, respectively.

We simulated derivation and validation samples of size N. For each subject we randomly generated a linear predictor: $\lambda_i \sim N(\mu = -0.84, \sigma^2 = 0.88)$. In doing so, we simulated data in which the distribution of the linear predictor was similar to that observed in heart failure patients in the EFFECT sample. In the derivation sample, the subject-specific probability of the outcome was $p_i = \frac{\exp(\lambda_i)}{1+\exp(\lambda_i)}$. However, in the validation or external sample, the subject-specific probability of the outcome was $p_i = \frac{\exp(k\lambda_i)}{1+\exp(k\lambda_i)}$. Thus, in the external validation sample, the effect of the linear predictor was shrunk by a factor of k. For each subject in the derivation and validation sample, a binary outcome was randomly generated from a Bernoulli distribution with parameter p_i.

In the derivation sample, a logistic regression model was used to regress the binary outcome on the simulated linear predictor λ. The estimated logistic regression model was then applied to the external validation sample to obtain a predicted probability of the occurrence of the outcome for each subject in the external validation sample. The simulations then proceeded as described previously with 500 iterations. Two factors were allowed to vary in a full factorial design: sample size (N = 500, 1000, and 10000) and the shrinkage factor (k = 0.6, 0.7, 0.8, 0.9, and 1.0). The last shrinkage factor (k = 1) denotes no shrinkage and serves as a control.

3.3.2. Over-fit model developed in derivation sample. We used the EFFECT-HF mortality prediction model to simulate data in which we examined the ability to detect external lack of calibration due to overfitting a regression model in the derivation sample. In the EFFECT sample, 2839 (33%) subjects died within 1 year of hospital admission. We standardized the four continuous predictor variables in the EFFECT-HF model (age, heart rate, respiratory rate, and urea level) so that each had mean zero and unit variance. We then determined the prevalence of each of the seven binary variables in the EFFECT-HF model (prevalences: 0.21, 0.17, 0.08, 0.17, 0.01, 0.12, and 0.13). We regressed the occurrence of death

within 1 year on the four standardized continuous predictor variables and the seven binary predictor variables in the EFFECT sample of patients hospitalized with heart failure.

We simulated data for both a derivation sample and an external validation sample. In each sample, we simulated four continuous predictor variables and seven binary predictor variables for each of N subjects. The four continuous predictor variables were sampled from independent standard normal distributions, while the seven binary predictor variables were sampled from independent Bernoulli distributions with parameters equal to the prevalences described in the previous paragraph. By using the regression coefficients estimated in the previous paragraph, a linear predictor and subject-specific probability of the outcome was determined for each subject in each of the derivation and validation samples. Note that the linear predictor was defined identically in the derivation and validation samples. We then fit a logistic regression model in the derivation sample in which the outcome was regressed on the 11 simulated predictor variables. The estimated logistic regression model was then applied to subjects in the external validation sample to obtain a predicted probability of the occurrence of the outcome. Internal and external calibration was assessed in the derivation and validation samples, respectively, by using previously described methods. In each simulated derivation sample, we determined the number of subjects who experienced the outcome along with the number of events per variable (EPV). The fitted loess curves and the number of events per variable were averaged over 500 iterations of the Monte Carlo simulations. We allowed the size of the simulated datasets to take on the following values: 200, 300, 500, 750, 1000, 2000, 5000, and 10,000.

4. Results – Monte Carlo simulations

4.1. Effect of c-statistic and incidence of the outcome on the assessment of calibration

The effect of the c-statistic of the underlying logistic regression model on the performance of the loess-based assessment of calibration is described in Figure 1. In each panel, we have plotted the estimated

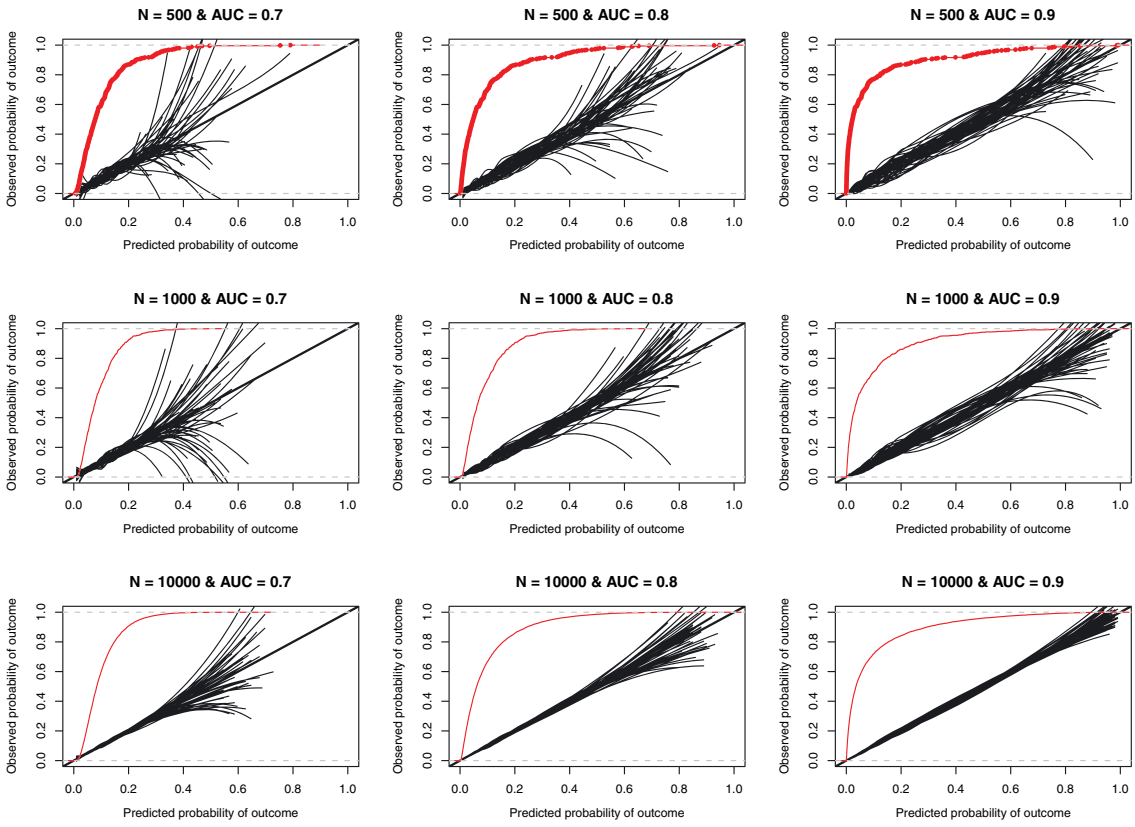

Figure 1. Effect of area under the receiver operating characteristic curve (AUC) and sample size on assessment of calibration.

loess model fit to each of the 50 simulated datasets. We have superimposed a diagonal line of slope 1 – this line depicts perfect calibration. Furthermore, on each panel we have also superimposed a red line depicting the empirical cumulative distribution function describing the distribution of the predicted probability of the occurrence of the outcome in one of the 50 simulated datasets (we elected to use the empirical cumulative distribution function because, similar to the observed vs. predicted plots, it is restricted to the unit square. Other methods such as nonparametric density estimates would not result in figures that were restricted to the unit square, thus necessitating a change in scale of the plots that could make details harder to discern).

For a given sample size, the degree of variation between the fitted loess curves decreased as the c-statistic increased. Similarly, for a given c-statistic, the degree of variation decreased within increasing sample size. The fitted loess lines displayed very good calibration and negligible variability within the range of predicted probability in which the majority of subjects lay. As the c-statistic of the logistic regression model used in the data-generating process increased, there was an increase in the width of the range of predicted probability in which the majority of subjects lay. Within a given scenario, there was modest to substantial variation across the fitted loess curves in the extreme upper tail of the distribution of the predicted probability of the occurrence of the outcome. However, this between-curve variation decreased as the c-statistic increased. It also decreased with increasing sample size. Thus, a key finding is that with higher discrimination, we can better assess calibration.

The effect of the incidence of the outcome on the performance of the loess-based assessment of calibration is described in Figure 2. As in the prior set of simulations, the fitted loess lines displayed very good calibration and negligible variability within the range of predicted probability in which the majority of the subjects lay when the incidence of the outcome was either 0.1 or 0.5. When the incidence of the outcome was 0.01 and the sample size was small to moderate (500 or 1000), it was relatively difficult to assess calibration. As the incidence of the outcome approached 0.5, there was an increase in the width of the range of predicted probabilities in which the majority of subjects lay. Furthermore, across scenarios, the degree of variation in the fitted loess curves in the extreme upper tail of the distribution of the predicted probability of the occurrence of the outcome decreased as the incidence of the outcome

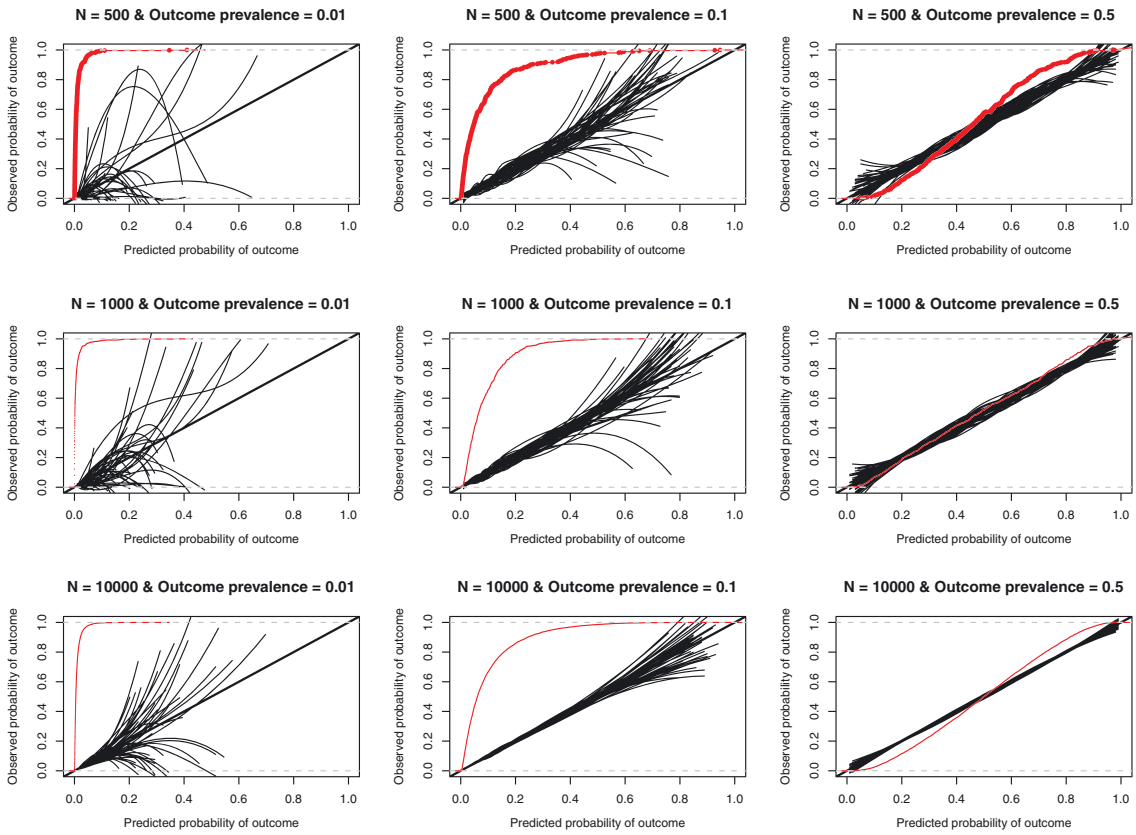

Figure 2. Effect of outcome prevalence and sample size on the assessment of calibration.

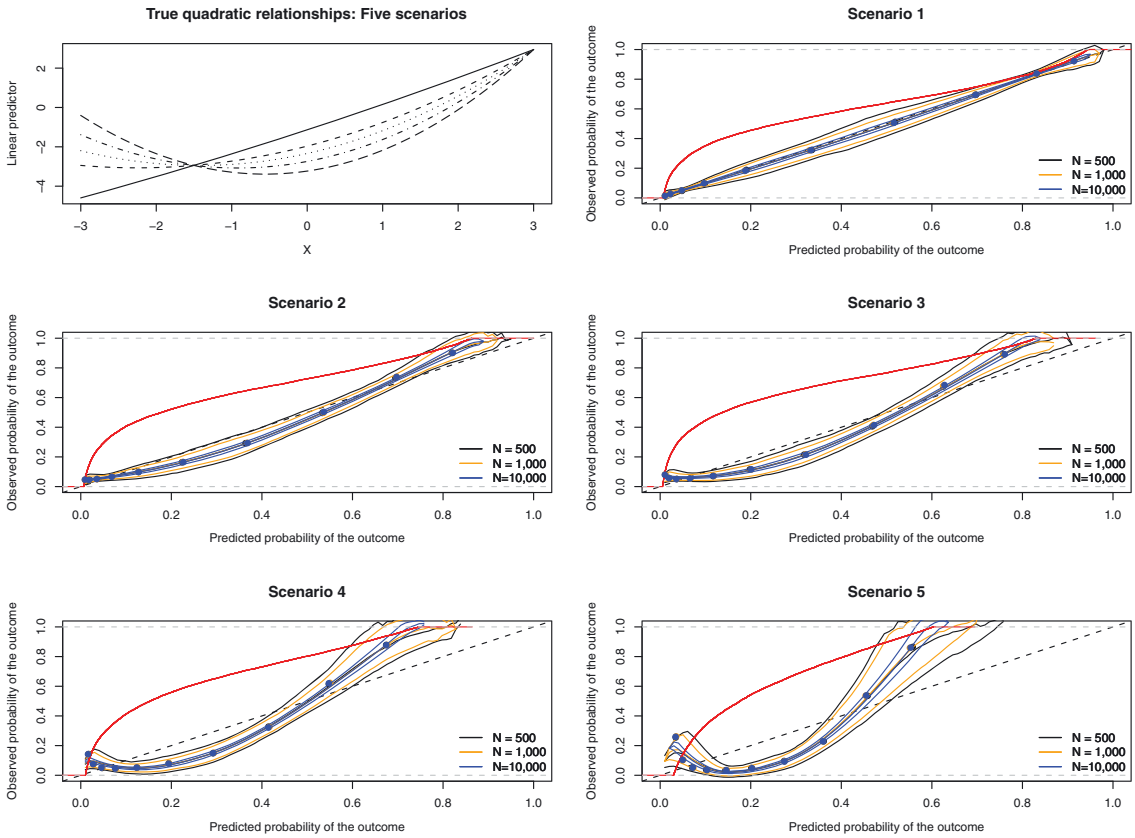

Figure 3. Quadratic relationship.

approached 0.50. Furthermore, this variation in the extreme upper tail of the distribution of the predicted probability of the occurrence of the outcome decreased with increasing sample size.

4.2. Assessing internal calibration or model misspecification

The ability of loess-based graphical assessment of calibration to detect the omission of a quadratic term from the fitted logistic regression model is described in Figure 3. The top-left panel describes the relationship between the continuous predictor variable and the log-odds of the outcome in the five scenarios. While the degree of nonlinearity increased across the five scenarios, the first scenario displayed almost no nonlinearity.

In the first scenario, with the least departure from linearity, the loess curves were not able to identify that the fitted model had been misspecified. However, in the remaining four scenarios, in which there was greater departure from linearity, the loess curves indicated the lack of internal calibration, providing evidence that the regression model had been misspecified. The loess curves displayed increasing nonlinearity as the true regression model exhibited increasing nonlinearity. In the presence of moderate to strong nonlinearity, the loess curves were, on average, able to detect model misspecification regardless of the size of the simulated datasets.

The results for the comparing observed versus predicted probability of the occurrence of the outcome across the ten deciles of risk are reported in Figure 4. In this figure, we report, for each risk stratum, the mean observed probability of the outcome across the 500 simulated datasets along with the mean of the mean predicted probability of the outcome. We also report estimated 95% confidence ellipses for these two means. In comparing Figures 3 and 4, one observes that the mean observed and mean predicted probabilities across the 500 simulated datasets produces the same qualitative interpretation as the mean loess-curves in Figure 3. However, the use of the ten strata of risk resulted in estimates that display greater variability than is evident in the loess-based method (i.e., the 95% confidence ellipsoids are wider than the 95% confidence intervals around the loess curves).

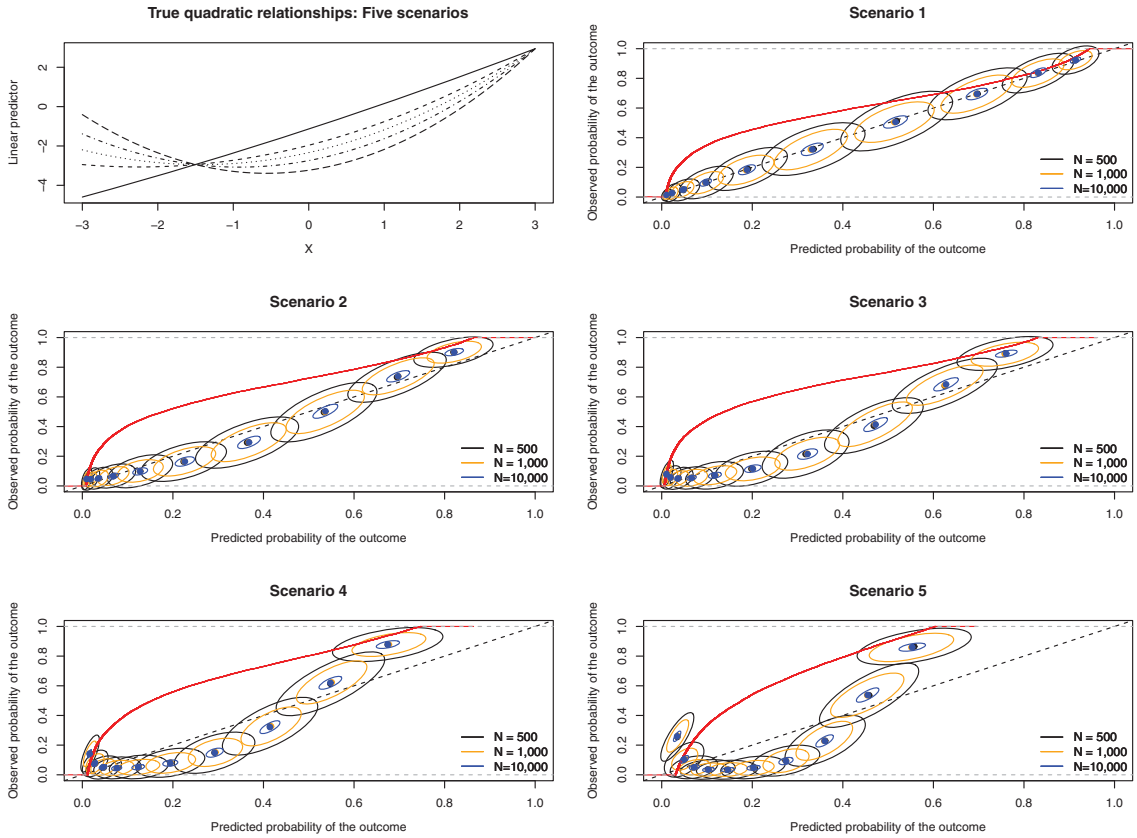

Figure 4. Quadratic relationship.

For each of the five scenarios, we determined the mean c-statistic of the correctly specified logistic regression model and the c-statistic of the incorrectly specified model across the 500 datasets of size 10,000. The pairs of c-statistics were (0.901, 0.901), (0.881, 0.881), (0.871, 0.865), (0.864, 0.835), and (0.868, 0.775) across the five scenarios. In the first two scenarios, there was no change in c-statistic due to fitting a misspecified regression model, while the decrease was 0.093 in the fifth scenario.

The ability of loess functions to detect the omission of both a binary covariate and its interaction with a continuous covariate from the fitted logistic regression model is described in Figure 5. The top-left panel describes the degree of interaction in each of the four scenarios. The lowest of the five straight lines describes the linear relationship between the continuous predictor variable and the log-odds of the outcome amongst subjects with $d = 0$. The upper four lines describe this relationship amongst subjects with $d = 1$ in the four different scenarios. In all of the four scenarios examined, on average, the loess curves did not provide any evidence of the lack of internal calibration, regardless of the size of the simulated datasets.

The ability of loess functions to detect the omission of the interaction between a binary covariate and a continuous covariate from the fitted logistic regression model is described in Figure 6. When the degree of interaction was strong, the loess curves provided modest graphical evidence of the lack of internal calibration. However, when the magnitude of the interaction was weak or moderate, then the loess curves were not, on average, able to provide evidence of model misspecification.

The results for the comparing observed versus predicted probability of the occurrence of the outcome across the ten deciles of risk are reported in Figures 7 and 8. As with the omission of a quadratic term, the use of the ten strata of risk resulted in estimates that displayed greater variability than was evident in the loess-based method.

For each of the four scenarios, we determined the mean c-statistic of the three different regression models across the 500 datasets of size 10,000. The triplets of c-statistics were (0.601, 0.599, 0.590), (0.667, 0.665, 0.628), (0.729, 0.727, 0.661), and (0.802, 0.800, 0.691) across the five scenarios (where each triplet consists of the c-statistic of the correctly specified model, the c-statistic of the model with

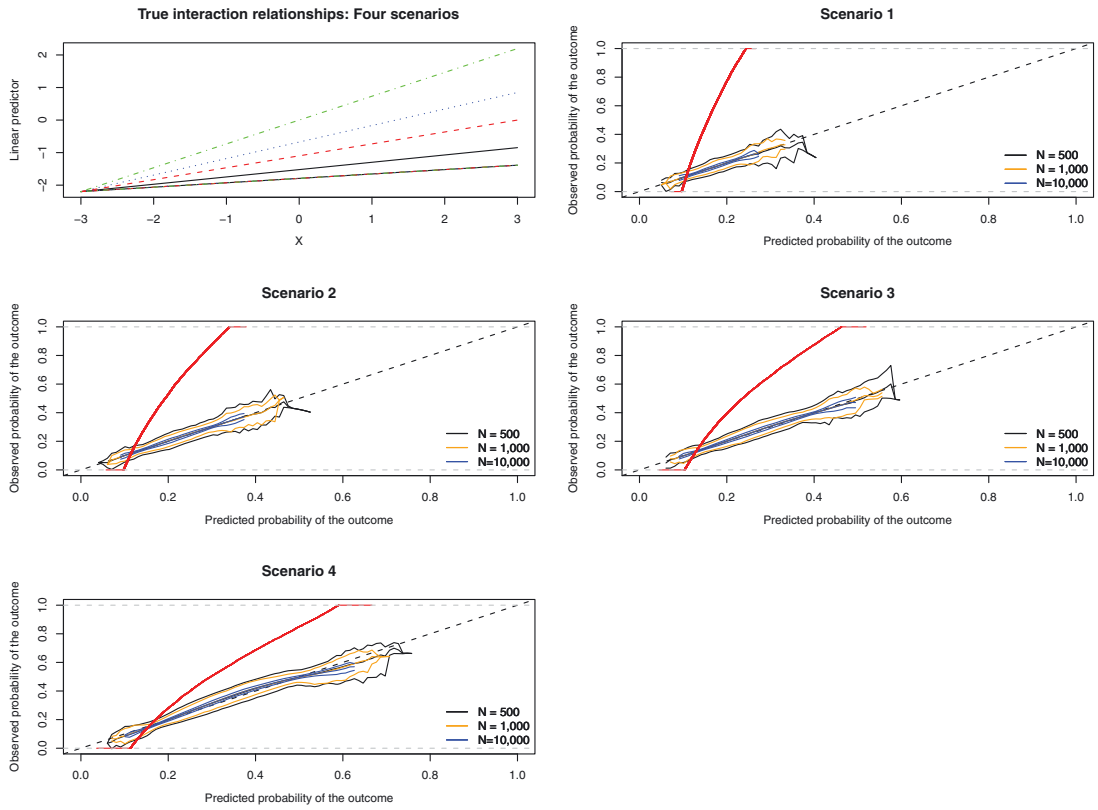

Figure 5. Interaction relationship: omitted binary variable and interaction term.

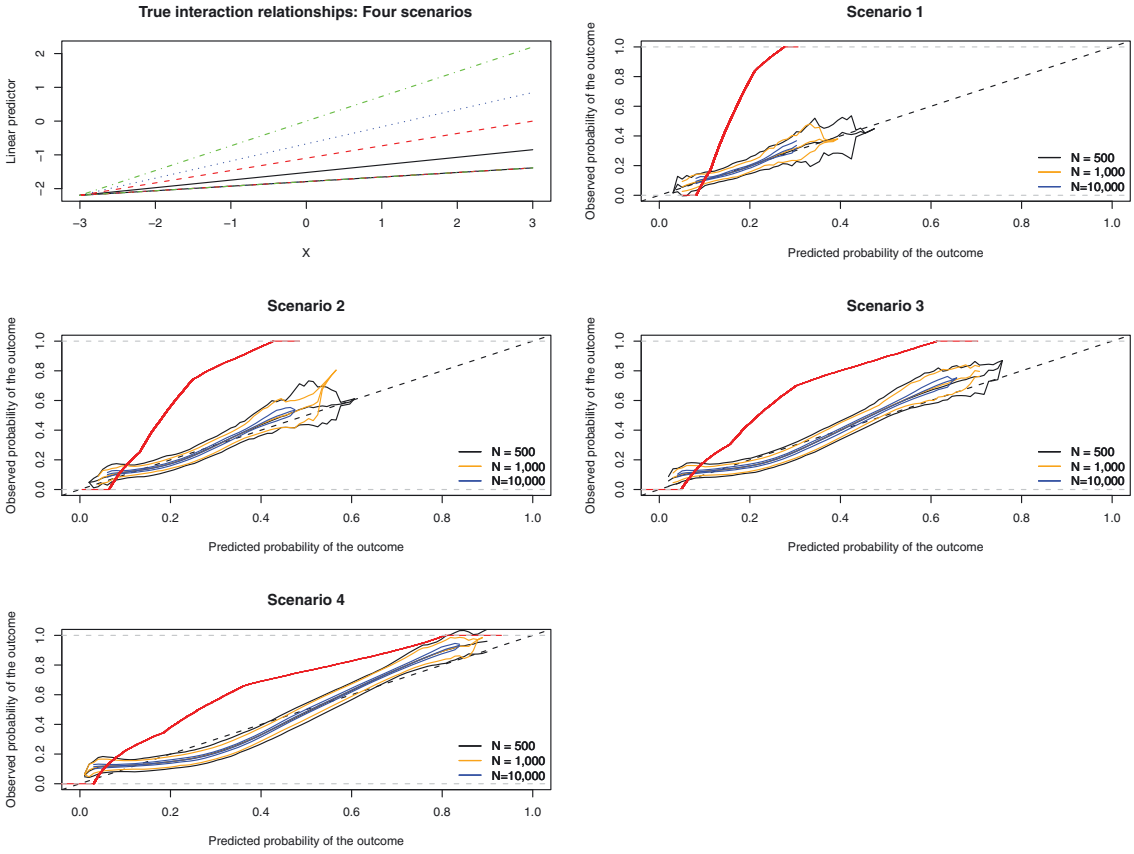

Figure 6. Interaction relationship: omitted interaction term.

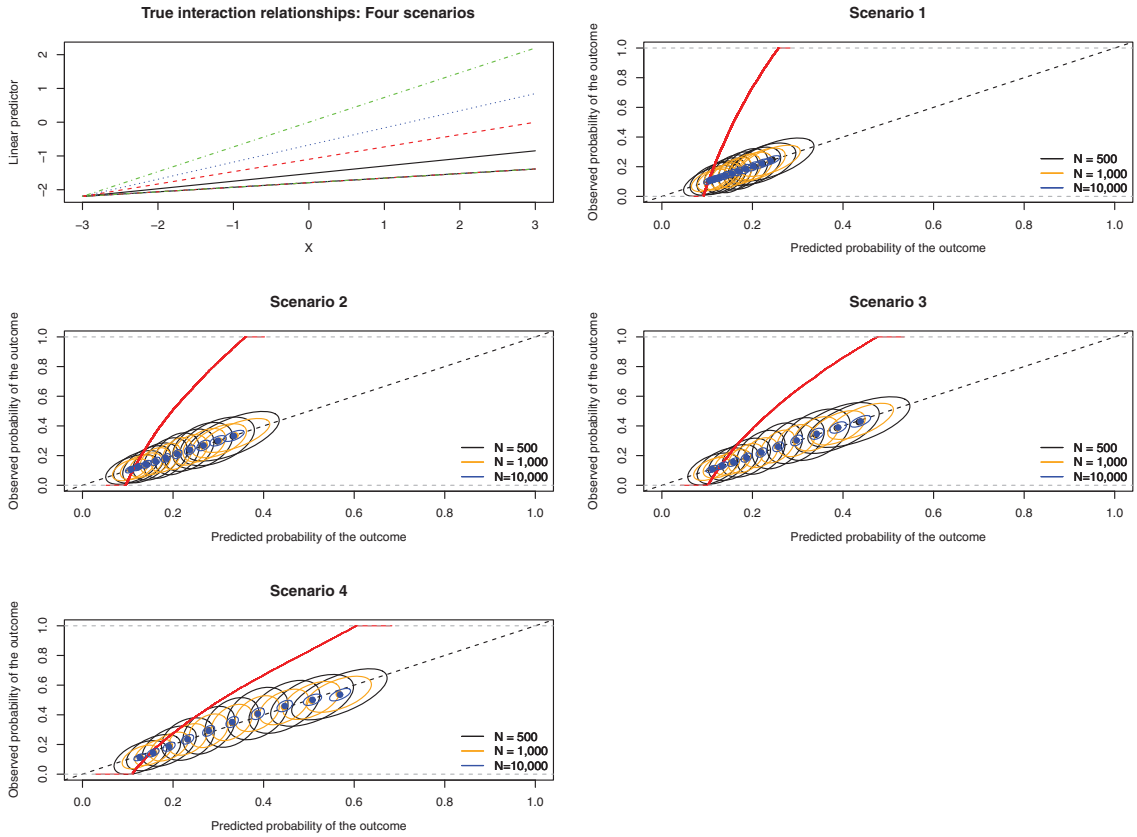

Figure 7. Interaction relationship: omitted binary variable and interaction term.

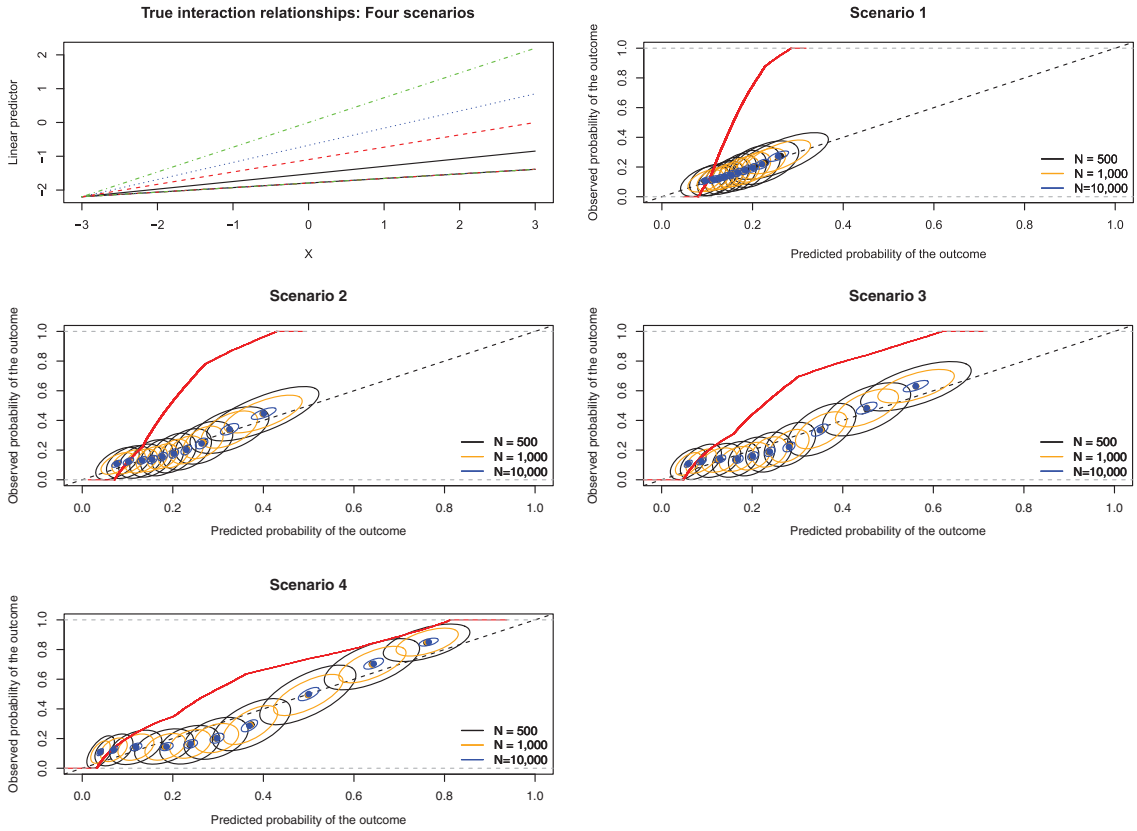

Figure 8. Interaction relationship: omitted interaction term.

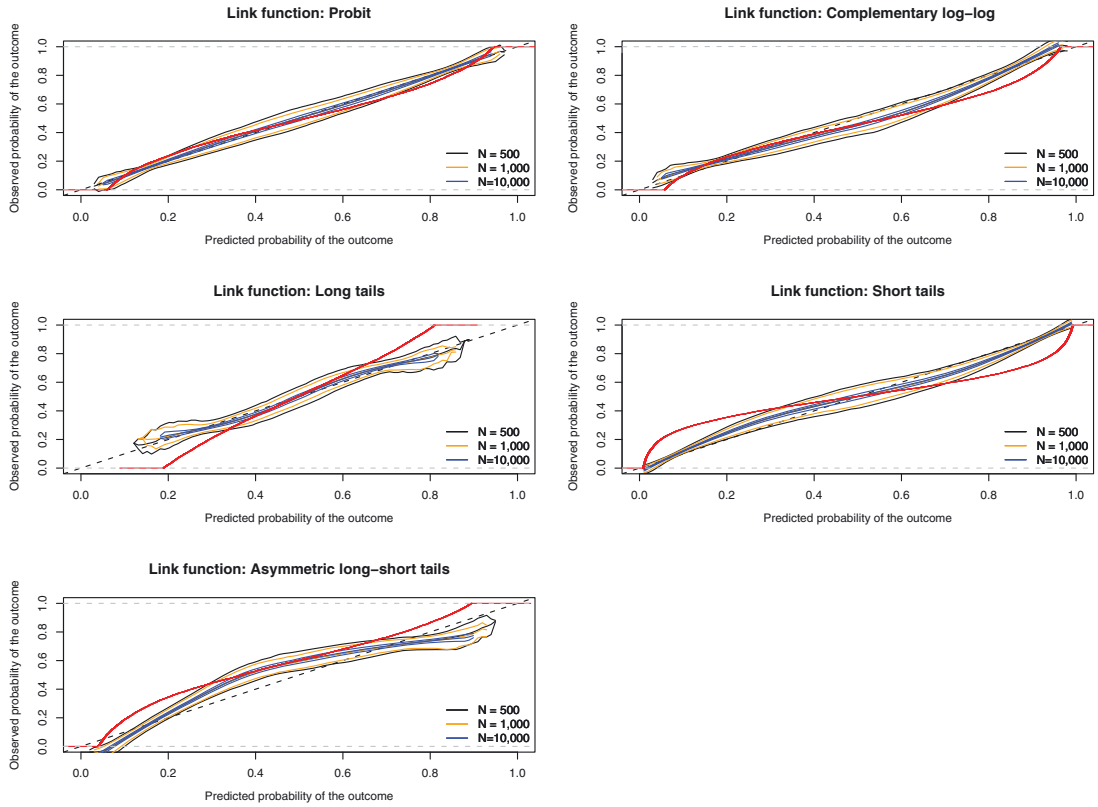

Figure 9. Different link functions.

the interaction omitted, and the c-statistic of the model with both the interaction and the binary covariate omitted). Omitting the interaction between the continuous and binary covariate resulted in a negligible change in the c-statistic. Omitting both the interaction and the binary covariate resulted in a greater change in the c-statistic, and the change in c-statistic increased across the four scenarios.

The ability of loess functions to detect misspecification of the link function is described in Figure 9. When the generalized logistic function had almost the same shape as the probit distribution or had longer tails than the logistic distribution, then the loess curves did not provide graphical evidence of the lack of internal calibration or of misspecification of the logistic regression model. When the generalized logistic function has approximately the same shape as the complimentary log-log function or had shorter tails than the logistic distribution, then the loess curves provided very modest graphical evidence of the lack of internal calibration. When the generalized logistic function was asymmetric with long and short tails, then there was moderately strong evidence of the lack of internal calibration.

The results for the comparing observed versus predicted probabilities of the occurrence of the outcome across the ten deciles of risk are reported in Figure 10. As with the omission of a quadratic term or the omission of an interaction, the use of the ten strata of risk resulted in estimates that display greater variability than was evident in the loess-based method.

Misspecification of the link function would not result in a change in the model c-statistic because each link function is a monotone transformation of the probability of the outcome. Thus, the rank ordering of predicted probabilities would not change.

4.3. Assessing external calibration

Graphical analyses of lack of external calibration due to a shrunken linear predictor in the external validation sample are described in Figure 11. When the degree of shrinkage was large ($k = 0.6$ or 0.7), loess-based methods were well able to detect lack of calibration in the external validation sample. However, when the degree of shrinkage was small ($k = 0.9$), it was very difficult to detect lack of external calibration graphically.

The ability of loess-based methods to detect lack of external calibration due to an over-fit model is described in the right panel of Figure 12. The left panel of this figure serves as a negative control: it

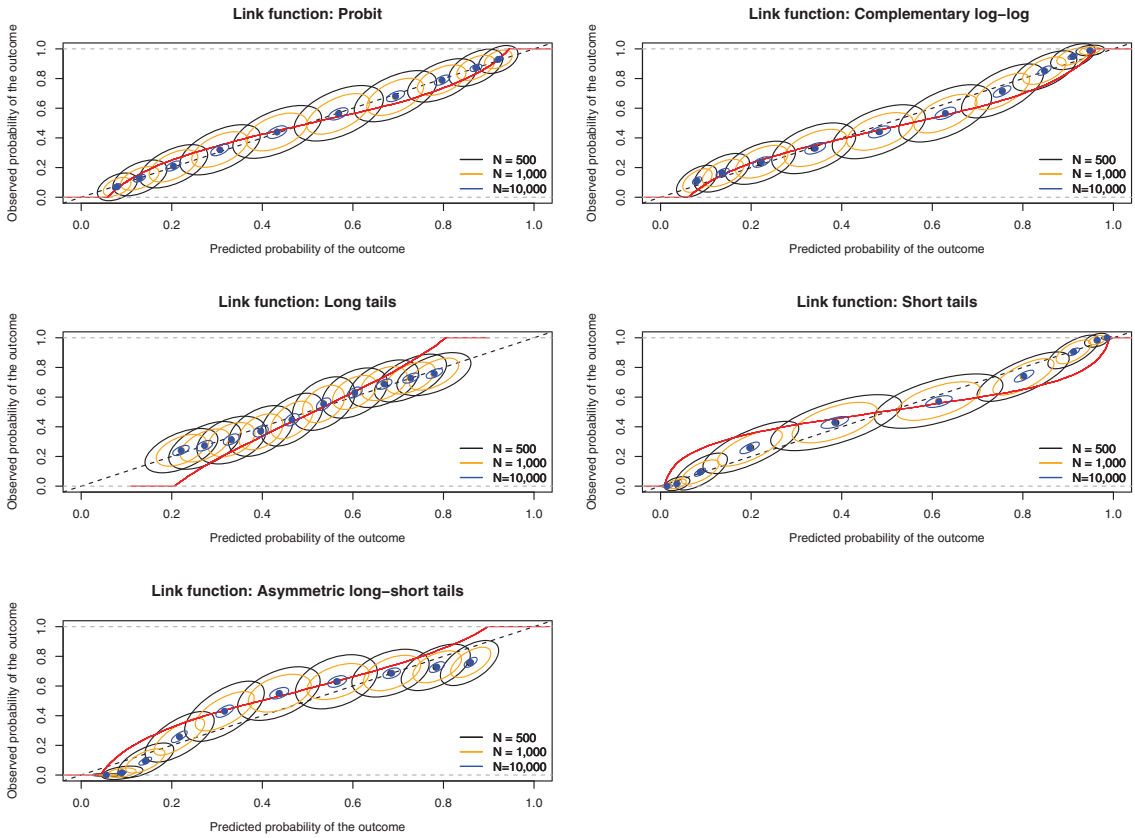

Figure 10. Different link functions.

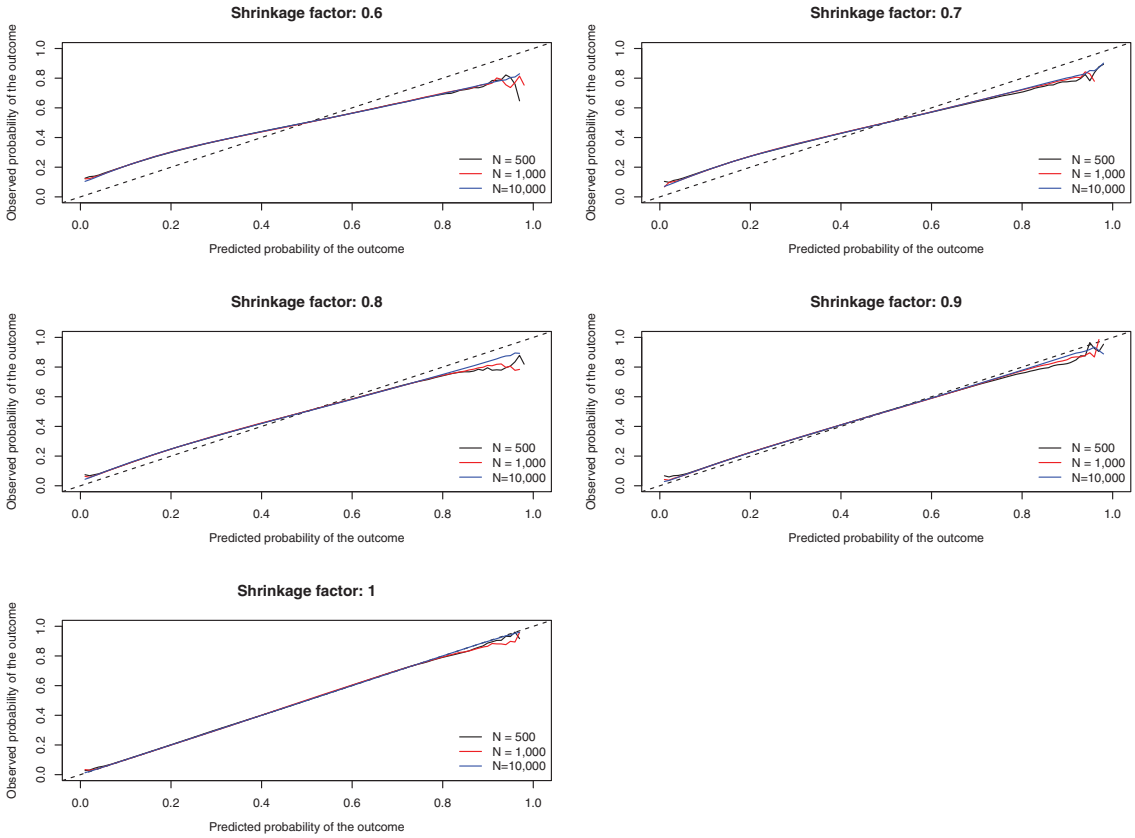

Figure 11. Shrunken linear predictor in validation sample.

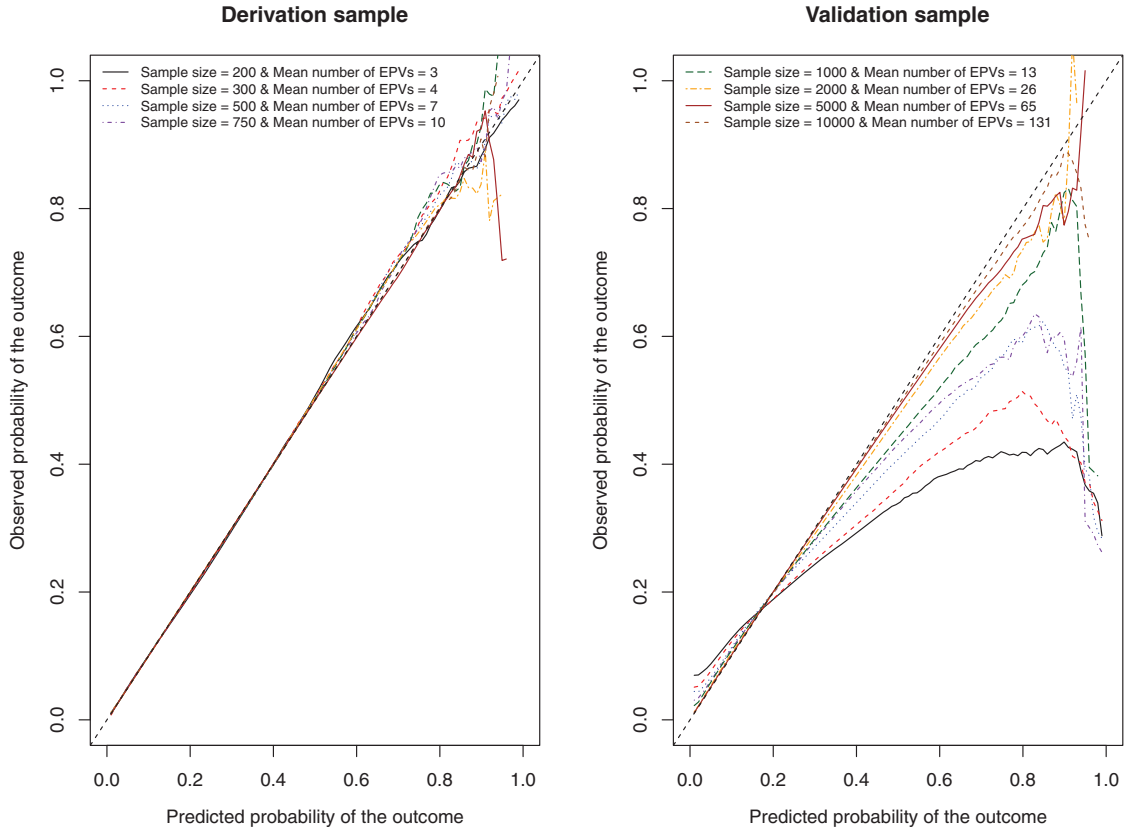

Figure 12. Calibration of overfit prediction model in derivation/validation samples.

examines the lack of internal calibration in the derivation sample in which the regression model was initially estimated. In the left panel, we observe that even when the number of EPV was very low, there was very little graphical evidence of lack of internal calibration. Any graphical evidence of lack of calibration was only apparent in the upper range of predicted probability and for a very low number of EPV. In stark contrast to this, there was strong evidence of lack of external calibration, even when the number of EPV was high. Even when the number of EPV was 13, there was still strong graphical evidence of the lack of external calibration.

5. Discussion

The assessment of the calibration of a model or algorithm for estimating the probability of the occurrence of a binary outcome is a key step in the validation of prediction models. Different authors have suggested that calibration be assessed using smoothed regression models to examine the relationship between the observed probability of the occurrence of the outcome and the predicted probability of the outcome. We conducted an extensive set of Monte Carlo simulations to examine the ability of a locally weighted least squares regression smoother to assess both internal and external calibration. We summarize our findings in the following two paragraphs and then place our findings in the context of the existing literature.

One set of findings relates to the ability of loess-based methods to detect model misspecification (lack of internal calibration). If we miss moderate to strong nonlinearity for a continuous predictor variable, we can graphically detect the lack of calibration. Similarly, if the fitted regression model omitted an interaction between a continuous predictor variable and a binary predictor variable, and the magnitude of the interaction was moderate to strong, then the loess-based methods were able to provide moderate evidence of model mis-specification. But even if the true link function was asymmetric with long-short tails, then these graphical methods were able to provide only modest evidence of the lack of calibration when a logistic link function was used in the fitted model. A second set of findings relates to the ability of loess-based graphical methods to detect poor calibration due to overfitting (lack of external calibration).

The ability to detect lack of external calibration due to a shrunken linear predictor was good when the degree of shrinkage was high, but it diminished as the degree of shrinkage decreased.

Secondary analyses also provided useful information for interpreting graphical calibration curves in specific settings. As expected, we observed that the utility of these curves tended to increase with increasing sample size, as there was decreased variability in the fitted curves. But even with small sample sizes, the loess algorithm behaved well, with smooth calibration curves. Furthermore, as the c-statistic of the underlying logistic regression model increased, the proportion of subjects who were in the extreme upper tail of the distribution of predicted probability increased. Consequently, there was diminished variability in the fitted loess curves. As a result, internal calibration could be more accurately assessed. Thus, a key finding is the relationship between discrimination and calibration: higher discrimination enables a better assessment of calibration. Finally, as the incidence of the outcome approached 0.50, the range of predicted probability increased. This resulted in decreased variability in the fitted loess curves, allowing for a better assessment of internal calibration. Synthesizing these two observations, would suggest that loess-based methods will perform better in settings in which the outcome has an incidence that is closer to 0.5 and in which the discrimination of the underlying model is good to excellent, rather than in settings in which the outcome is rare (or extremely common) and the discrimination of the underlying model is relatively poor.

Several studies have examined the statistical power of different tests of calibration or goodness-of-fit for the logistic regression model. One of the most extensive of these was by Hosmer *et al*. [24]. Several of our simulations considered scenarios identical to those described in their paper, so that our results could be compared directly with their results. They demonstrated that the statistical power to detect the omission of a binary covariate and its interaction with a continuous covariate was very low when the sample size was 500 (power less than 12% across all five scenarios and across nine different goodness-of-fit tests). Our findings in this instance were similar: the use of graphical methods of assessing calibration did not provide evidence of lack of calibration. Thus, both formal statistical testing and graphical assessment had limited ability to detect the omission of a binary covariate and its interaction. However, when considering the regression model in which only the interaction term was omitted, we found that if the magnitude of the interaction was moderate to strong, then loess-based methods were able to provide modest to strong evidence of lack of calibration. When examining departures from nonlinearity, Hosmer *et al*. found that, in the two scenarios with the strongest nonlinearity, power was uniformly very high to detect lack of fit when the sample size was 500. Similarly, we found that loess-based methods were able to provide evidence of lack of calibration and that the strength of the graphical evidence increased with increasing nonlinearity.

We conducted a limited examination of the practice of dividing subjects into strata of risk according to the deciles of the predicted probability of the occurrence of the outcome. The observed versus mean predicted probability of the outcome was then compared within each stratum. While, on average, this method performed similarly to the loess-based methods, the variability in observed versus mean predicted probabilities was greater than the observed variability in the fitted loess curves.

In examining graphical methods for calibration, we have made a distinction between assessment of calibration in the dataset in which the model was developed (internal calibration) and assessment of calibration in external validation datasets (i.e., samples other than that in which the regression model was developed). When developing prediction models, the focus is frequently on the latter. When the developed regression model will be applied to new patients or populations other than the one in which it was developed, model developers indeed should strive for a model in which predictions are well calibrated in these external populations. Thus, a focus on external calibration takes precedence over internal calibration when the aim is to predict for new subjects outside the specific derivation sample.

However, if the model is being developed primarily for the purpose of understanding patterns of predictor effects within a specific sample, then a focus on internal calibration is warranted. Also, internal calibration can be of central importance to hospital profiling, in which observed mortality is compared with the expected mortality at each hospital within a jurisdiction or health care network. Iezzoni suggested that if the purpose of the regression model is to compare expected mortality at a given hospital (as derived from the fitted regression model) to observed mortality at that hospital, then the assessment of calibration should supersede that of the assessment of discrimination [29]. Indeed, recent work emphasizes that the c-statistic is irrelevant in judging the quality of an adjustment model for hospital report cards, regardless of whether one is using a conventional logistic regression model to compare observed to expected mortality or whether one adds hospital-specific random effects to the conventional logistic regression model [30].

The assessment of model fit in the derivation sample can directly be determined using standard tests during model derivation [22]. For instance, to test the assumption that a continuous variable is linearly related to the log-odds of the outcome one could use a likelihood ratio test to compare a regression model that assumed a linear effect of the given predictor variable with a regression model that used restricted cubic splines functions to model the given relationship [31]. Similarly, statistical hypothesis testing can be used to test whether an interaction is statistically significant and warrants retention in the final model. Our findings indicate that graphical methods for assessing model calibration in the derivation sample may serve a complementary adjuvant role for these purposes. One might consider that the use of graphical methods for assessing internal calibration has the advantage in that it may free the analyst from conducting multiple statistical tests to assess the presence of interactions. Especially in a setting with a large number of predictor variables, standard hypothesis testing raises the risk of including spurious interactions in the model. Instead, graphical assessment of calibration provides a global assessment of the adequacy of the fitted model in the sample in which it was derived, similar to using overall tests of interaction [8]. Graphical assessment allows one to focus on developing a model that predicts accurately, despite possibly have minor errors in specification.

There are limitations to the use of graphical methods to assess calibration. In particular, the interpretation of the deviations from the line of identity is, to an extent, subjective. Different analysts or readers may apply different criteria as to what constitutes meaningful lack of calibration. In contrast, test-based approaches to detect model fit appear to offer greater objectivity to assessing calibration. However, it should be noted that the application of methods such as the Hosmer–Lemeshow test requires decisions about the number of risk strata into which the sample is divided. Furthermore, the Hosmer–Lemeshow test is directly influenced by sample size. Statistically significant lack-of-fit will often be detected in large datasets [32]. This can make it difficult to compare the calibration of different models in datasets of different size. Recently, Paul, Pennell and Lemeshow developed recommendations for the number of groups into which to divide the sample so that the power of the Hosmer-Lemeshow test could be standardized across datasets of different sizes [32]. In addition, they made the following recommendation '*For samples larger than* $n = 25,000$, *we do not recommend the use of the Hosmer-Lemeshow test*' (page 75). The graphical methods that we have examined do not suffer from these limitations of test-based methods. In particular, graphical methods should perform very well in large samples, the exact setting in which the use of the Hosmer–Lemeshow test has been discouraged.

Our findings also provide insight into a debate about the relative merits of discrimination and calibration. Diamond argued that a prediction model cannot be both perfectly calibrated and perfectly discriminatory [33]. Furthermore, he suggested that a 'model that maximizes discrimination does so at the expense of reliability (calibration)' (page 88). It should be stressed that his derivations were based on the, rather artificial, assumption that the probability of the occurrence of the outcome was uniformly distributed in the population. Our simulations show that this claim is not generalizable to other settings. In examining some of the panels in Figure 1, one observes than the regression model can have excellent discrimination (c-statistic of 0.90) and the fitted model can also have excellent calibration. Indeed, this is a very desirable property of a prediction model: for the model to discriminate well between those with and without the condition or outcome and for there to be good agreement between predicted and observed probabilities.

In summary, we found that loess-based methods perform well for assessing the calibration of logistic regression models. Their ability to graphically assess internal calibration and detect model misspecification was comparable to that of formal test-based approaches for assessing model fit. Furthermore, loess-based methods should be used for detecting the lack of external calibration due to the use of an overfit regression model.

Appendix A

Comparison of the loess and lowess functions in R and the effect of the span parameter on assessment of internal calibration.

A.1. Introduction

In the simulations examining the performance of loess-based methods for assessing calibration, we used the loess function in the R statistical programming language [25]. The key parameter in this function is the span parameter, which denotes the width of the window around each subject such that all subjects

within that window are used to fit the weighted least squares regression line used to obtain a prediction for a given subject [23]. In R, the lowess function implements an older implementation of the loess algorithm. The purposes of the limited set of analyses reported in this appendix were twofold: (i) to compare the performance of the loess function with that of the lowess function for assessing internal calibration; and (ii) to compare the impact of different choices of the span parameter on assessment of internal calibration.

A.2. Methods

We conducted a set of Monte Carlo simulations similar to those described in Section 3.1.1 of the paper (examining the effect of the c-statistic of the logistic regression model on the ability to assess calibration). We designed the data-generating process so that the outcome would occur for approximately 10% of subjects and so that the c-statistic of the logistic regression model would be approximately 0.80. We used a full factorial design in which we allowed two factors to vary. We considered three values of the span parameter: 0.25, 0.50, and 0.75 (the default in R). We considered three values of the sample size: 500, 1000, and 10,000. We thus considered nine different scenarios. The analyses that were conducted were identical to those described in Section 3.1.1 (with the exception that the span parameter of the regression smoother was allowed to vary from the default in R).

The following R code was used: loess($Y \sim P$, span $= k$) and lowess($Y \sim P$, f $= k$, iter $= 0$), where k denotes the value of the span parameter ($k = 0.25$, 0.50, and 0.75). The parameter denoted by iter in the lowess function is described as the number of 'robustifying' iterations. As noted by Harrell *et al.*, it is important that the value of this parameter be set to 0 [6].

A.3. Results

The effect of selecting different values for the span parameter of the loess function in R on the assessment of model calibration is described in Figure A.1 of the online Appendix. There are nine panels,

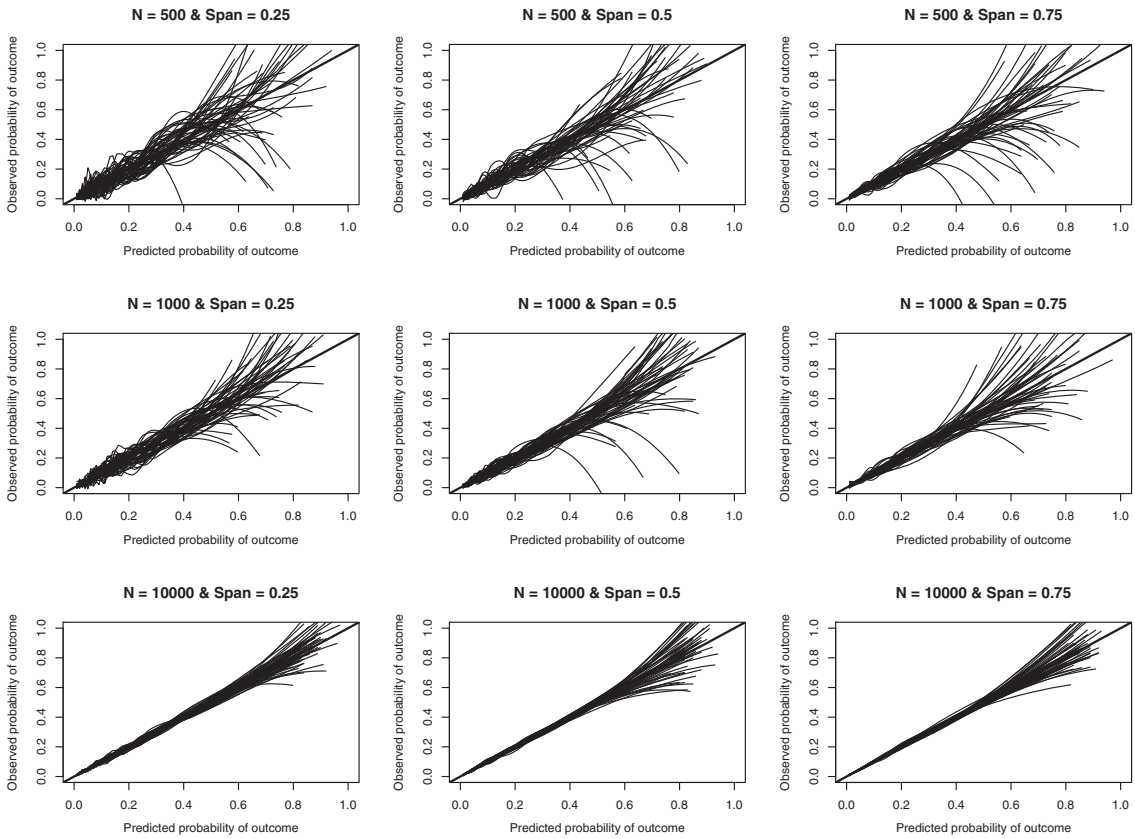

Figure A.1. Effect of span parameter and sample size on the assessment of calibration (loess function).

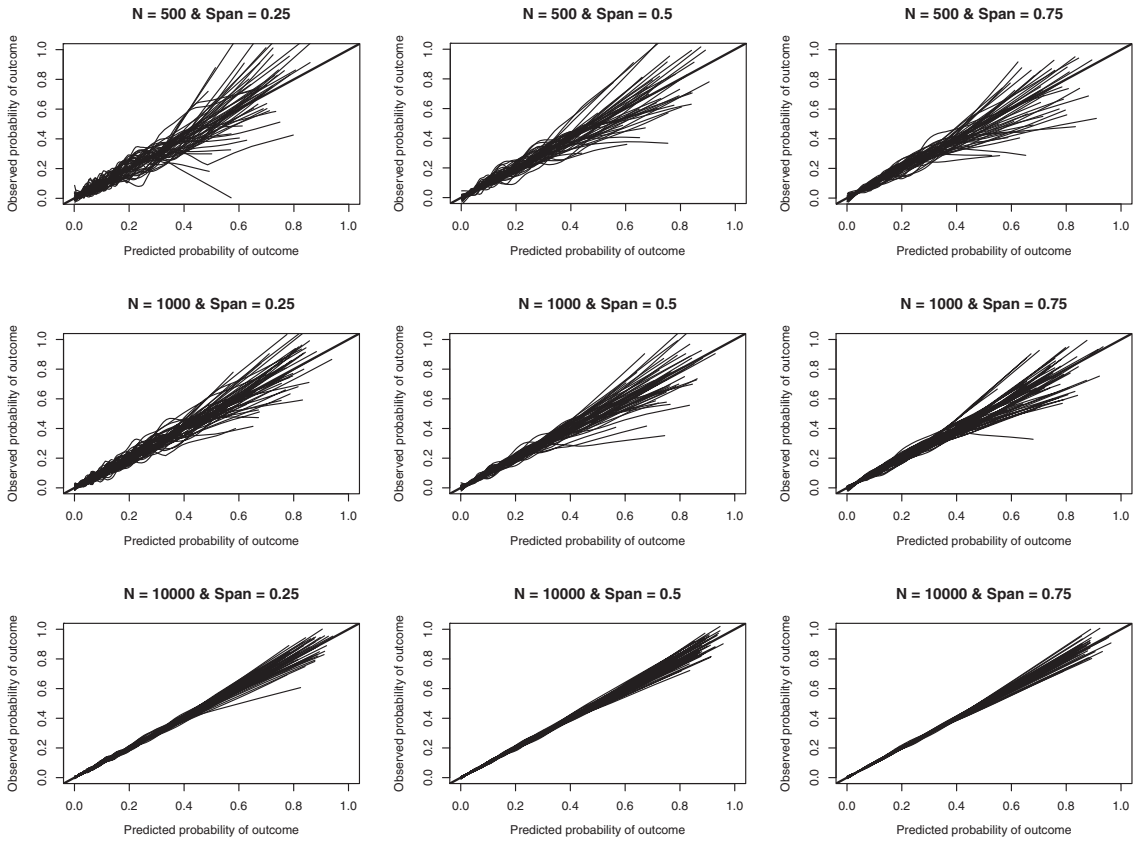

Figure A.2. Effect of span parameter and sample size on the assessment of calibration (lowess function)

one for each combination of the value of the span parameter and the sample size of the simulated dataset. Results using a given sample size are reported in the same row of the figure. In each panel, we have plotted the estimated loess model fit to each of the 50 simulated datasets. On each panel, we have superimposed a diagonal line of slope one – this line depicts perfect calibration.

For a given value of the span parameter, variation between the fitted loess curves decreased as the sample size increased. When the sample size was large ($N = 10,000$), there was very little variation between the fitted loess curves within the range of predicted probability in which the majority of subjects lay. However, there was moderate variation between fitted curves in the extreme upper tail of the distribution of predicted probability of the occurrence of the outcome. For a given sample size, variation between fitted loess curves decreased as the degree of smoothing increased (i.e., as the span parameter increased). However, the degree of decrease in variation diminished as the sample size increased. When the span parameter was set to the R default (0.75), there was relatively little variation between the fitted loess curves within the range of predicted probability in which the majority of the subjects lay. Because of the good performance of the default value in R (0.75), this value was used in all of simulations in the main body of the research body.

Similar results for the lowess function were observed (Figure A.2). In comparing Figures A.1 and A.2, one notes that the use of the lowess function resulted in modestly less variation in the smoothed regression curves in the upper tail of distribution of predicted probability.

Acknowledgements

This study was supported by the Institute for Clinical Evaluative Sciences, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care. The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (MOP 86508). Dr. Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation. Dr. Steyerberg is supported in part by the Netherlands Organization for Scientific Research (grant 9120.8004). The EFFECT data used in the study was funded by a

Canadian Institutes of Health Research Team Grant in Cardiovascular Outcomes Research. These datasets were held securely in a linked, de-identified form and analyzed at the ICES.

References

1. Breiman L. Random forests. *Machine Learning* 2001; **45**(1):5–32.
2. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer-Verlag: New York, NY, 2001.
3. Austin PC, Lee DS, Steyerberg EW, Tu JV. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? *Biometrical Journal* 2012; **54**(5):657–673. DOI: 10.1002/bimj.201100251.
4. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion). *The Annals of Statistics* 2000; **28**:337–407.
5. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology* 2013; **66**(4):398–407.
6. Harrell FE, Jr., 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**(4):361–387.
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**(1):128–138.
8. Steyerberg EW. *Clinical Prediction Models*. Springer-Verlag: New York, 2009.
9. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. John Wiley & Sons: New York, NY, 1989.
10. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics* 1980; **9**(10):1043–1069. DOI: 10.1080/03610928008827941.
11. Tsiatis AA. A note on a goodness-of-fit test for the logistic regression model. *Biometrika* 1980; **67**(1):250–251. DOI: 10.2307/2335347.
12. le Cessie S, van Houwelingen JC. A goodness-of-fit test for binary data based on smoothing residuals. *Biometrics* 1991; **47**:1267–1282.
13. Royston P. The use of customs and other techniques in modelling continuous covariates in logistic regression. *Statistics in Medicine* 1992; **11**(8):1115–1129.
14. Stukel TA. Generalized logistic models. *Journal of the American Statistical Association* 1988; **83**(402):426–431. DOI: 10.1080/01621459.1988.10478613.
15. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45**(3–4):592–565. DOI: 10.1093/biomet/45.3-4.562.
16. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stastisics in Medicine* 1991; **10**(8):1213–1226.
17. Harrell FE, Lee KL. Using Logistic Model Calibration to Assess the Quality of Probability Predictions, Division of Biometry, Duke University Medical Center. http://biostat.mc.vanderbilt.edu/wiki/pub/Main/FrankHarrell/logistCal.pdf.
18. Harrell FE, Jr. *Regression Modeling Strategies*. Springer-Verlag: New York, NY, 2001.
19. Dalton JE. Flexible recalibration of binary clinical prediction models. *Stastisics in Medicine* 2013; **32**(2):282–289.
20. Copas JB. Plotting p against x. *Applied Statistics* 1983; **32**(1):25–31.
21. Copas JB. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society - Series B* 1983; **45**(3):311–354.
22. Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 1984; **3**(2):143–152.
23. Cleveland WS, Grosse E, Shyu WM. Local Regression Models. In *Statistical Models in S*, Chambers JM, Hastie TJ (eds). Chapman & Hall: New York, 1993; 309–376.
24. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 1997; **16**(9):965–980.
25. R Core Development Team. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, 2005.
26. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC: Medical Research Methodology* 2012; **12**:82. DOI: 10.1186/1471-2288-12-82.
27. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *Journal of the American Medical Association* 2003; **290**(19):2581–2587.
28. Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, Ko DT. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *Journal of the American Medical Association* 2009; **302**(21):2330–2337.
29. Iezzoni LI. *Risk Adjustment for Measuring Health Outcomes (second edition)*. Health Administration Press: Chicago, 1997.
30. Austin PC, Reeves MJ. The relationship between the C-statistic of a risk-adjustment model and the accuracy of hospital report cards: a Monte Carlo study. *Medical Care* 2013; **51**(3):275–284. DOI: 10.1097/MLR.0b013e31827ff0dc.
31. Harrell FE, Jr., Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *Journal of the National Cancer Institute* 1988; **80**(15):1198–1202.
32. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Stastisics in Medicine* 2013; **32**(1):67–80.
33. Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *Journal of Clinical Epidemiology* 1992; **45**(1):85–89.

