Journal of Clinical Epidemiology 64 (2011) 993−1000

Journal of Clinical Epidemiology

Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure

Delphine S. Courvoisier[a,b,*], Christophe Combescure[a,b], Thomas Agoritsas[a,b], Angèle Gayet-Ageron[a,b], Thomas V. Perneger[a,b]

[a]Faculty of Medicine, University of Geneva, Switzerland

[b]Division of Clinical Epidemiology, University Hospitals of Geneva, Geneva, Switzerland

Accepted 24 November 2010

Abstract

Objective: Logistic regression is commonly used in health research, and it is important to be sure that the parameter estimates can be trusted. A common problem occurs when the outcome has few events; in such a case, parameter estimates may be biased or unreliable. This study examined the relation between correctness of estimation and several data characteristics: number of events per variable (EPV), number of predictors, percentage of predictors that are highly correlated, percentage of predictors that were non-null, size of regression coefficients, and size of correlations.

Study Design: Simulation studies.

Results: In many situations, logistic regression modeling may pose substantial problems even if the number of EPV exceeds 10. Moreover, the number of EPV is not the only element that impacts on the correctness of parameter estimation. High regression coefficients and high correlations between the predictors may cause large problems in the estimation process. Finally, power is generally very low, even at 20 EPV.

Conclusion: There is no single rule based on EPV that would guarantee an accurate estimation of logistic regression parameters. Instead, the number of predictors, probable size of the regression coefficients based on previous literature, and correlations among the predictors must be taken into account as guidelines to determine the necessary sample size. © 2011 Elsevier Inc. All rights reserved.

Keywords: Model adequacy; Model building; Type I error; Power; Event per variable; Logistic regression

1. Introduction

Logistic regression modeling is commonly used in health research, either to identify independent risk factors for health outcomes or to build diagnostic and prognostic models. This type of modeling does not guarantee accurate results [1]. A common problem occurs when the outcome has few events with respect to the number of candidate predictors; then the estimated odds ratios (ORs) can be biased, and the final model may be overfitted to the development sample [2,3]. As yet, there is no consensus on the number of events needed per variable [3]. On theoretical grounds, Harrell et al. [4,5] proposed that 10−20 events per variable (EPV) were necessary. A simulation study based on the real data set by Peduzzi et al. [6] concluded that 5−10 EPV were enough. Vittinghoff and McCulloch [7] conducted a broader set of simulations and showed that in most cases, confidence interval coverage and relative bias were appropriate with a minimum EPV of five. Although these results are reassuring, the general validity of the rule of thumb of 5, 10, or even 20 EPV requires testing in a broader set of situations or data structures, with a broader set of evaluation criteria.

We sought to extend previous work on the relation between EPV and the performance of logistic regression modeling, by varying not only EPV but also the correlations between predictors, strengths of ORs, number of predictors, and frequency of non-null associations. We also examined five evaluation criteria—frequency of nonconvergence of regression models, relative bias of the area under the receiver operating characteristic curve (AUC), relative bias of regression coefficients, confidence interval coverage, and power. Also, we examined three typical situations—univariate analyses (one predictor); multiple regression conducted with a large set of potential predictors (we chose 25), some of which may have no association with the

* Corresponding author. Division of Clinical Epidemiology, University Hospital of Geneva (HUG), 6, rue Perret-Gentil, 1205 Geneva, Switzerland. Tel.: +41-22-3729029; fax: +41-22-3729135.

E-mail address: delphine.courvoisier@hcuge.ch (D.S. Courvoisier).

0895-4356/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2010.11.012

What is new?

- In many situations, logistic regression modeling may pose substantial problems even if the number of events per variable (EPV) exceeds 10.
- Additionally to the number of EPV, high regression coefficients and high correlations between the predictors may cause problems in the estimation process.
- Power is generally very low even at 20 EPV.
- To determine the sample size needed to obtain accurate logistic regression parameters, the number of EPV, probable size of the regression coefficients based on previous literature, and correlations among the predictors must be taken into account.

outcome; and multiple regression conducted with a smaller set of non-null predictors (we chose seven), a situation that may arise in a later stage of modeling, after covariates that show weak associations with the outcome have been eliminated from consideration.

2. Method

2.1. Simulation scheme

We conducted a factorial simulation study of logistic regression with continuous multivariate normal predictors of mean 0 and variance 1. We considered values of EPV of 3, 5, 7, 10, 15, 20, and 25; models with 1, 7, or 25 predictors; and values of regression coefficient parameters of log(1), log(1.2), log(1.5), log(2), and log(3). For models with several predictors, the correlations between predictors were 0.2, 0.5, or 0.7. A proportion of the predictors (28%—2/7 for 7 and 7/25 for 25 predictors, 56% [4/7 and 14/25], or 100%) had non-null regression coefficients, whereas the others were null. Similarly, only a proportion of the predictors (28%, 56%, or 100%) were highly intercorrelated, whereas the others were correlated at 0.2. Note that ORs of 2 per standard deviation for continuous predictors and correlations of 0.7 between variables are rare; but these designs were included to thoroughly examine model performance in all situations. For all simulations, the proportion of events was around 50%. To examine the influence of the proportion of events, we also simulated univariate models with only 10% of events by manipulating the intercept.

For each design of the simulation study, a population data set ($N = 100,000$) was generated based on the number of predictors, correlations between predictors, and regression coefficients using the software program R version 2.9.2. [8]. Five hundred replications (random samples) were then drawn from this population. For each replication, observations were drawn in sequence until the desired number of EPV was met. Logistic regression of the population data set provided the true parameters and AUC, and logistic regression of the samples yielded the estimates.

2.2. Model performance

Four main problems may occur when estimating parameters using logistic regression models. First, the estimation may not converge; second, the estimation may be biased; third, the estimation may be imprecise and vary across samples, which will impact whether the confidence interval contains the population value (coverage); and fourth, the standard error may be under- or overestimated, leading to type I or II errors. To examine the accuracy of model-based standard errors, we also compared the average of standard errors obtained over all samples with the standard deviation across samples (results not shown because the average standard errors and standard deviation were always very close for all simulation designs). Thus, model performance was estimated by four indices (denoted as a–d). Note that all indices are summarized by the median and not by the mean because the distributions of the indices are not normal.

a. **Percentage of nonconverged replications**
Nonconvergence mainly occurs when there is no overlap, that is, when knowing the values of the predictors allows a perfect prediction of the event. However, an overlap of only a few observations often leads to extreme parameters estimates. When a parameter estimate corresponded to an OR of 50 or more (or 1/50 or less), we considered that the replication did not converge. Other indicators are based only on the replications that converged.

b. **Relative bias of regression coefficients**
Median differences between the population parameter and each replication's parameter divided by the true parameter value. For better readability of the results, when there were several predictors with identical OR, we computed the median of the bias over all these predictors. Relative bias >0.15 may be considered problematic [7].

c. **Confidence interval coverage**
Percentage of replication in which the confidence interval includes the true parameter value. It should be equal to 95%. We computed the median of the confidence interval coverage over all predictors with identical OR. Values <0.93 were considered problematic [7].

d. **Percentage of statistically significant logistic regression coefficients at level 0.05**
It is an indicator of power when the true parameter is non-null and of type I error when the true parameter is null (OR = 1). We computed the median of the percentage of significant coefficients over all predictors with identical ORs.

Finally, to examine the problems that may occur when using logistic regression for prediction purposes, we also present the relative bias of the AUC (relative bias of the

AUC is the median of differences between the true [population] AUC and estimated AUC divided by the true AUC).

2.3. Results presentation

We first present the results of univariate analyses, in which only regression coefficients and EPV influence the estimation of the parameter. We then present the results of the simulations of a large (25 predictors) multivariate analysis and of a smaller (seven predictors) multivariate analysis with non-null predictors. In these simulations, the size of the OR and EPV, as well as the correlation between predictors, percentage of predictors with high regression coefficients, and percentage of predictors that are highly correlated may influence parameter estimation.

3. Results

3.1. Univariate analysis

Fig. 1 presents the results of the simulations of univariate analyses.

a. **Percentage of nonconverged replications**
Less than 5% of the simulations do not converge for EPV greater than or equal to 10. When EPV is below seven, high OR is associated with an increasing amount of nonconvergence (up to 12% for three EPV).

b. **Bias**
The relative bias of the simulations in which the true regression parameter is non-null is close to 20% for all ORs when EPV is lower than or equal to 10.

c. **Confidence interval coverage**
Confidence interval coverage is close to its expected value (95%) for EPV equal to or higher than five.

d. **Percentage of statistically significant regression coefficients**
The type I error (ie, percentage of significant β when the OR = 1) is close to 0.01 for EPV smaller than 10. In the same conditions, the power (ie, percentage of significant β when the true OR is not 1) is also too low. Even 25 EPV are not enough to obtain a power above 80% for all ORs.

The results of the simulation with a proportion of 10% of events are similar to the results presented above, except

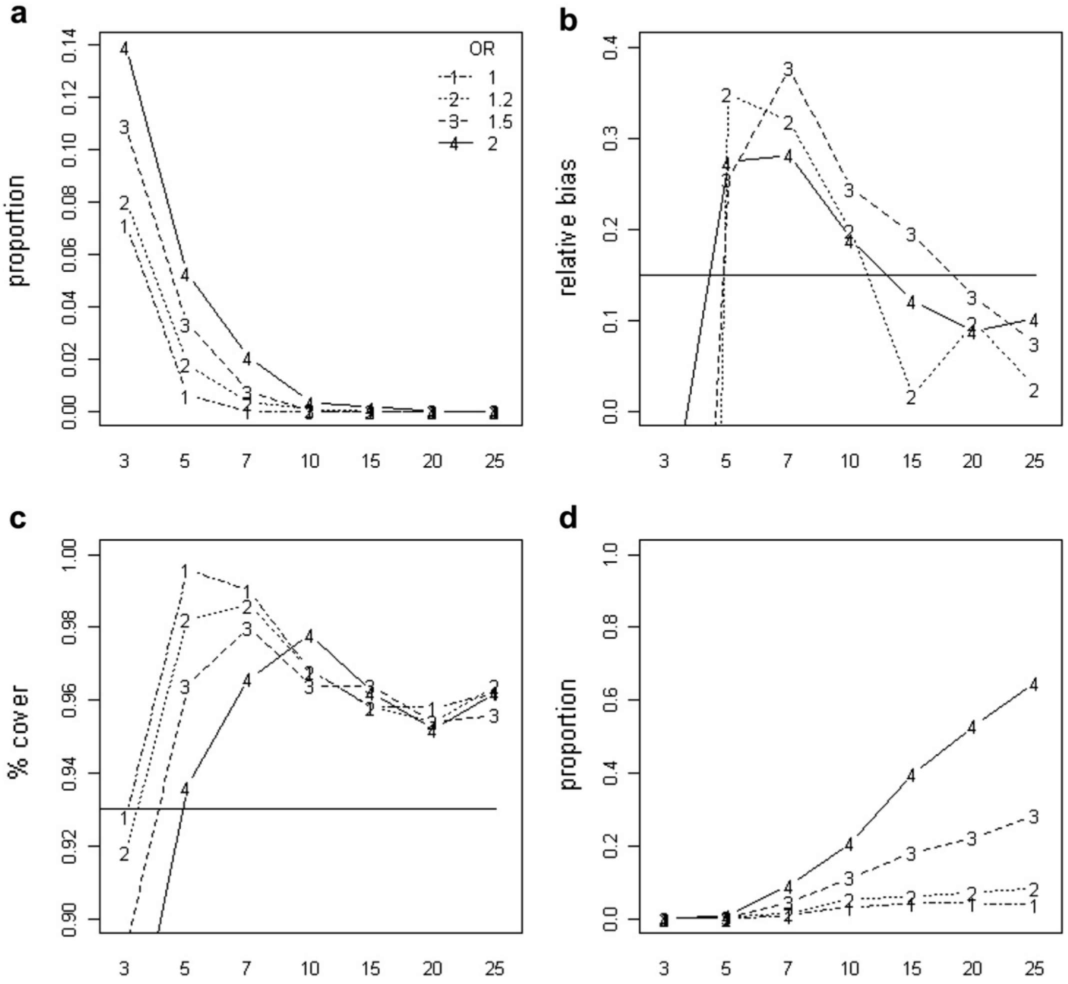

Fig. 1. Results of the simulation studies for the univariate analysis. X-axis indicates the number of events per variable. Horizontal lines indicate cutoffs proposed by Vittinghoff and McCulloch [7]. (a) Percentage of nonconverged replications. (b) Median relative bias of the estimate. (c) Percentage of cover of the confidence interval. (d) Percentage of significant coefficients. Inset is only presented in (a). OR, odds ratio.

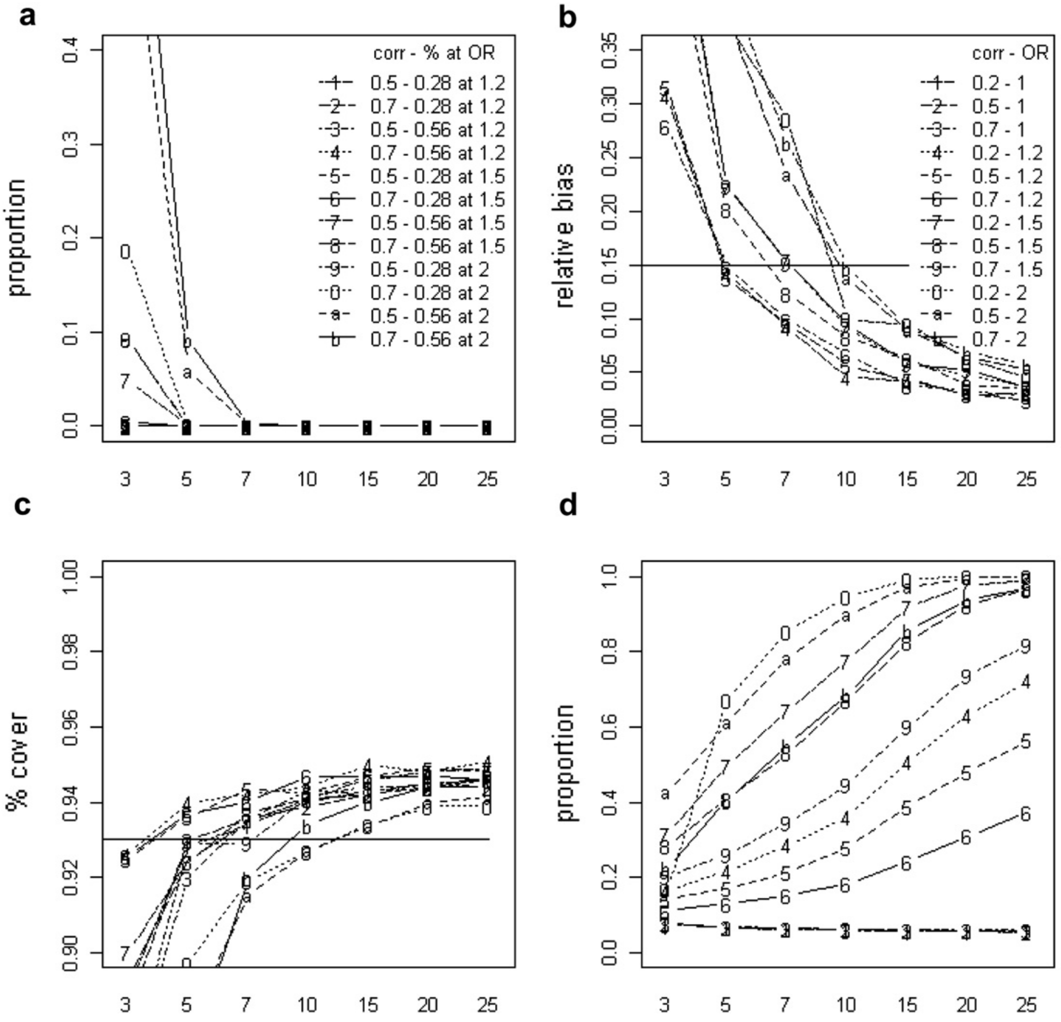

Fig. 2. Results of the simulations studies for the large multivariate analysis including 25 predictors. X-axis indicates the number of events per variables. Horizontal lines indicate cutoffs. (a) Percentage of nonconverged replications. (b) Median relative bias of the estimates. (c) Median % cover of the confidence interval. (d) Median % significant coefficients. Inset of (b) holds for (c) and (d). corr, correlation; OR, odds ratio.

that nonconvergence and relative bias of the coefficients decreases (data not shown). In contrast, the relative bias of the AUC remains high and the power stays low.

3.2. Multivariate analyses

Fig. 2 presents the results of the simulations of a large multivariate analysis including 25 predictors, without any selection procedure. Each model performance indicator is shown for simulation of varying designs (2 × 2 × 3 designs): the predictors could be correlated at 0.5 or 0.7, and 28% or 56% of the predictors could be non-null with ORs values of 1.2, 1.5, or 2.

1. **Percentage of nonconverged replications**
   The size of the true regression coefficients has the largest impact on nonconvergence. The percentage of non-null predictors also matters, with a larger percentage corresponding to lower convergence. More specifically, replications with low ORs (1.2 or 1.5; lines 1–8) almost always converge for EPV equal to five or more. However, when the ORs are equal to two and the percentage is high (lines a and b), the percentage of nonconverged replications rises close to 10% for five EPV.
2. **Bias**
   The relative bias of non-null regression coefficients is equal to or lower than 0.15 for EPV equal to or higher than 15. However, when EPV equal 10, high OR (lines 0, a, and b), irrespective of predictors' correlations, leads to a median relative bias close to 0.15. Relative bias then increases rapidly until, at five EPV, all simulations except those with OR equal to 1.2 are heavily biased. The absolute bias of the simulations in which regression coefficients equal 0 is close to zero independently of the number of EPV (data not shown).
3. **Confidence interval coverage**
   For OR of 2 (lines 0, a, and b), confidence interval coverage is acceptable for 10 EPV. For low ORs (1, 1.2, and 1.5), even seven EPV yield correct coverage.

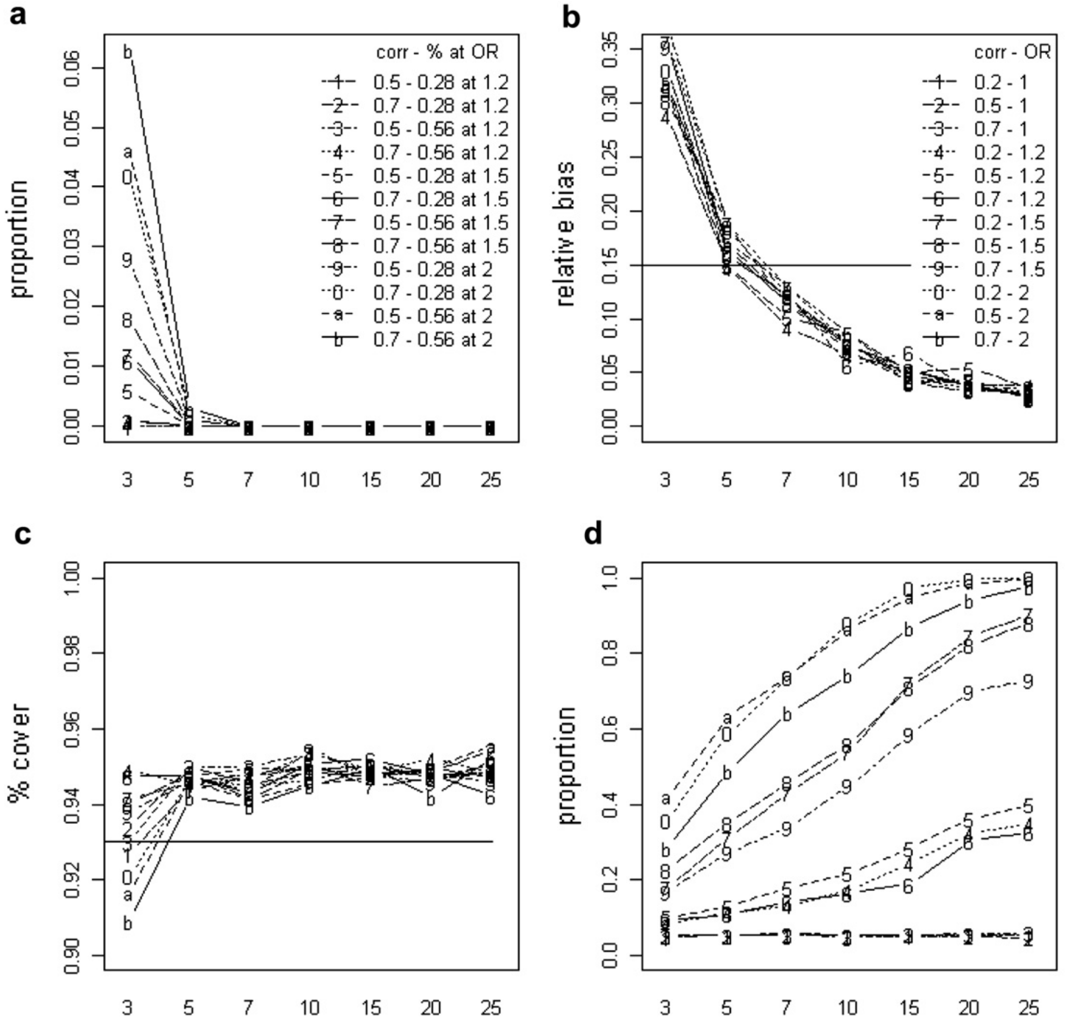

Fig. 3. Results of the simulation studies for the parsimonious multivariate analysis including seven homogeneous predictors. X-axis indicates the number of events per variables. Horizontal lines indicate cutoffs. (a) Percentage of nonconverged replications. (b) Median relative bias of the estimates. (c) Median % cover of the confidence interval. (d) Median % significant coefficients. Inset of (b) holds for (c) and (d). corr, correlation; OR, odds ratio.

4. **Percentage of significant regression coefficients**
   The type I error is appropriate for all EPV. On the other hand, the power is too low. As expected, power increases with the OR and number of EPV. However, power decreases when the correlations between predictors increase. In particular, even 25 EPV are not enough to obtain a power above 80% for an OR of 1.2 (lines 4–6). At 10 EPV, OR of 1.5, irrespective of predictors' correlations, has less than 80% power (lines 7–9). Even an OR of 2, when the predictors are highly correlated, does not reach 80% power (line b). Finally, EPV of seven or less almost never yield 80% power.

The results of the simulations of a smaller (seven predictors) multivariate analysis with non-null predictors are presented in Fig. 3, when 28% or 56% of predictors are correlated, and in Fig. 4, when all predictors are correlated. Null predictors were not included because, in clinical research, variables in the final model often have at least some explanatory power.

1. **Percentage of nonconverged replications**
   Nonconvergence is very low except when all predictors are correlated. In that case, it increases as regression coefficients and correlation between predictors increase. However, at seven EPV, all simulations still lead to less than 10% of nonconvergence.
2. **Bias**
   When all predictors are correlated, the relative bias for the predictors is equal to or lower than 0.15 for EPV of seven or more and the data structure has almost no influence. Only the number of EPV determines the strength of the bias. When the percentage of correlated predictors is low, the relative bias for the predictors is lower than 0.15 for EPV of 15 or more. However, when EPV equal 10, high ORs, especially if the predictors are moderately (0.5) or highly (0.7) correlated (lines 8 and 9), lead to a median relative bias close to 0.15. Finally, at five EPV, all simulations except one (OR = 1.2, correlations between predictors = 0.2) are heavily biased.

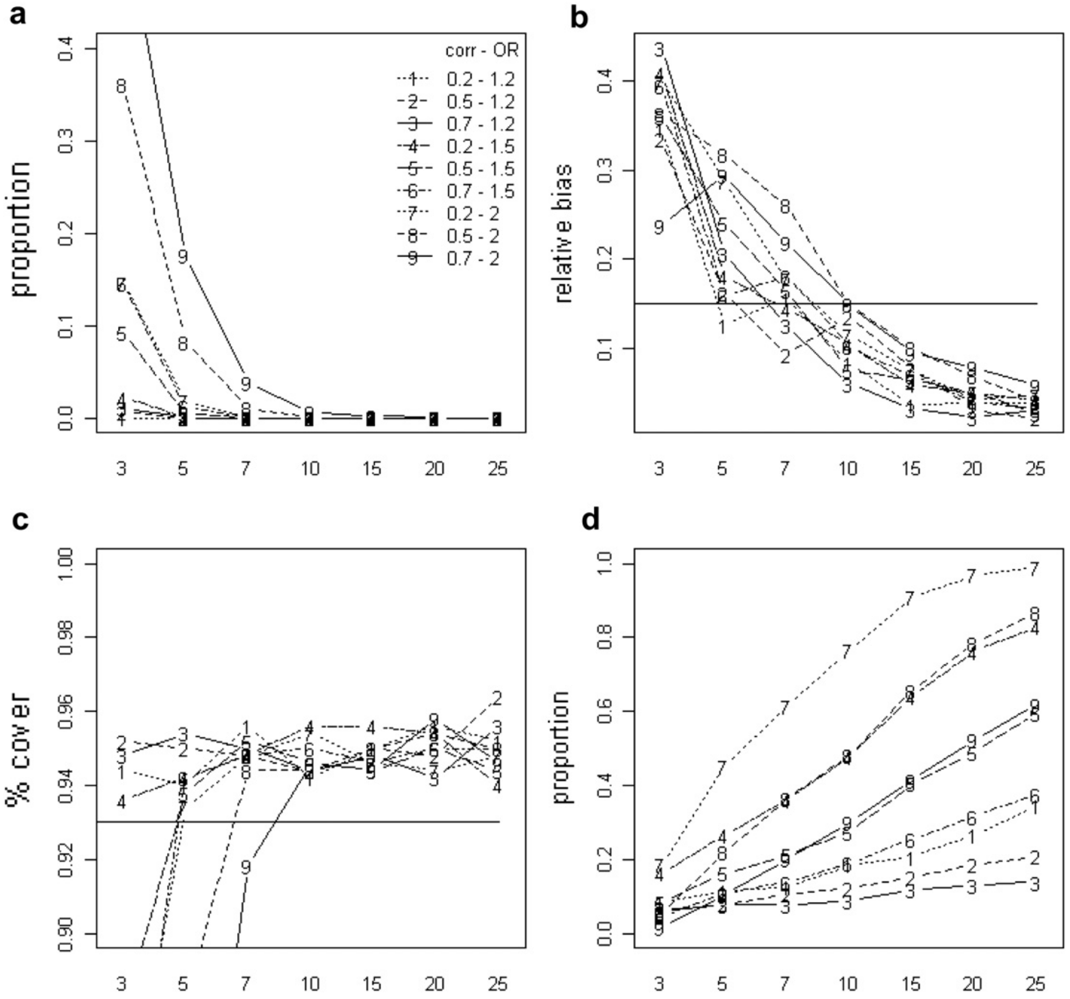

Fig. 4. Results of the simulation studies for the parsimonious multivariate analysis including seven nonhomogeneous predictors. X-axis indicates the number of events per variables. Horizontal lines indicate cutoffs. (a) Percentage of nonconverged replications. (b) Median relative bias of the estimates. (c) Median % cover of the confidence interval. (d) Median % significant coefficients. Inset is only presented in (a). corr, correlation; OR, odds ratio.

3. **Confidence interval coverage**
   At 10 EPV, all confidence interval coverages are acceptable. However, at five EPV, the confidence interval coverage when all predictors are correlated with OR = 2 (lines 8 and 9) is too low.
4. **Percentage of significant regression coefficients**
   Similarly to the previous simulations, power increases with the OR and EPV and when the correlations between predictors decrease. Nonetheless, power is generally low. Indeed, even at 15 EPV, ORs below 2 do not yield a power of 80%.

3.3. Area under the receiver operating characteristic curve

When the relationship between the predictor(s) and the outcome is weak, the AUC is overestimated, especially when there are few predictors (Fig. 5). For instance, when the OR of a single predictor is equal to 1 or 1.2, the relative bias in the AUC is higher than 10% at 10 EPV. On the contrary, when there are many predictors (25), the relative bias in the AUC is almost always lower than 10%. When there are seven predictors, the relative bias of the AUC increases for low OR and a small percentage of non-null coefficients and rises to more than 10% at five EPV when OR = 1.2.

4. Discussion

These simulation studies show that the number of EPV is not the only element that influences the correctness of the predictive capacity of the model and of parameter estimation in logistic regression models when all predictors are of interest. The sizes of the parameters and correlations between the predictors also play an important part in the estimation process. In univariate analyses, 10 EPV lead to median relative biases for the predictors above cutoff (0.15) and low power (<80%), even for high regression coefficients. In large multivariate analyses (25 predictors), the proportion of nonconverged replications can be very high, especially when a high proportion of the predictors are

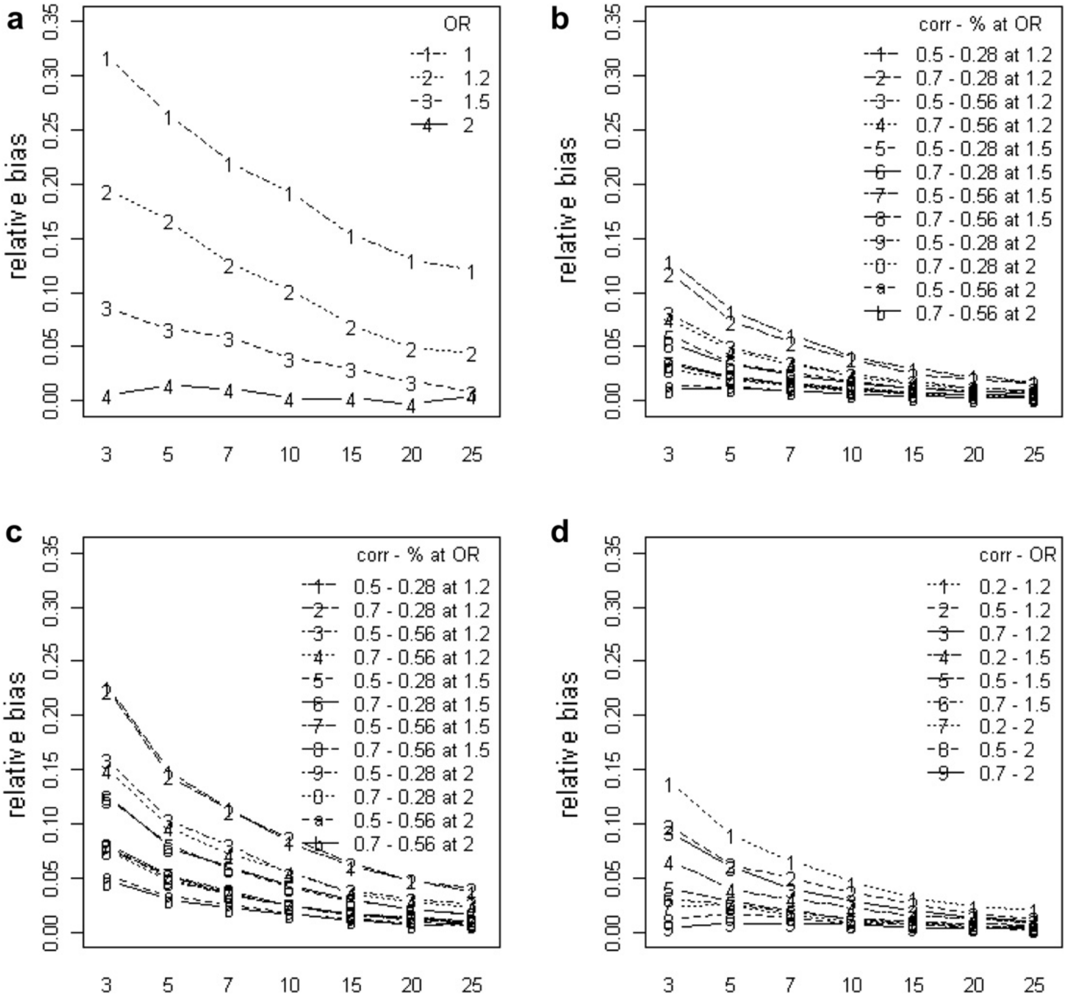

Fig. 5. Relative bias of the area under the receiver operating characteristic curve. X-axis indicates the number of events per variables. (a) One predictor. (b) Twenty-five predictors. (c) Seven predictors. (d) Seven predictors: all correlated. corr, correlation; OR, odds ratio.

non-null. High regression coefficients and high correlations between predictors lead to high relative bias for the predictors even at 10 EPV, which may seem counterintuitive because necessary sample size is usually smaller when expected effect is higher. This is because logistic regression is of limited use when the outcome can be almost perfectly predicted by the independent variables. Moreover, high correlations between predictors diminish power. In smaller multivariate models (seven predictors), the results are similar and the loss of power is as problematic, but the degree of bias is not as important. These results moderate previous conclusions [6,7] that 10 EPV or even less are enough to obtain good parameter estimation. Finally, the AUC of a model with weak low predictive power (ie, few predictors with that are low OR weakly associated with the outcome) is often overestimated.

Two results are noteworthy and unexpected. First, the AUC is biased when the number of predictors and number of EPV are low and the model is not very predictive (population AUC close to 0.5 and true OR close to 1). In the univariate case, when the true parameter is null, the estimated coefficient b will be negative half the time and positive half the time. Accordingly, AUC(X) will be less than 0.5 half the time and greater than 0.5 the other half. Not so, however, for the predicted probability and its estimated AUC(bX). When b is positive, AUC(bX) equals AUC(X). But when b is negative, AUC(bX) equals AUC($-$X) or, equivalently, 1 $-$ AUC(X). Thus the estimated AUC(bX) will always be greater than 0.5, and the expectation of this statistic will be biased vis-à-vis the true population value of 0.5. This explains why the AUC of non-predictive models will frequently be overestimated in small samples. Others have shown that the AUC estimated on a first sample is often higher than it is on the subsequent samples [9,10]. The artifact found in this study may contribute to this overestimation in AUC.

The second noteworthy result is that the power of the regression models was often very low, even for 20 or 25 EPV when the OR is below 2 [11]. For comparative studies, the computation of sample size to achieve the desired power is well codified. In contrast, no such standard procedure exists for model building and multivariate analyses. A commonly used rule of thumb is that the number of EPV should be greater than 5 or 10. This rule of thumb will lead to insufficient power, so that variables that actually predict the outcome will be found nonsignificant in the initial model and, in the case of model building, dropped from the prognostic model. Thus, data structure should always be taken into account to obtain an estimate of necessary sample size (eg, Ref. [12]).

Taken together, the results of this study imply that researchers should explore the correlations of their predictors of interest and should be careful about including several highly correlated predictors into a logistic regression model. Possible solutions to this problem include the selection of uncorrelated predictors based on clinical criteria or the computation of a single score representing all correlated predictors (eg, through factor analysis).

One limitation of this study is that the indices of good estimation were not presented for each parameter in each simulation design. Instead, the median of these indices was taken over all predictors of similar regression coefficients, and the medians of these medians were taken over all simulations with a similar design. This synthetic presentation of results will inevitably hide outliers, which may be interesting in their own right. Another limitation is that we did not explore all possible data structures and hence were not able to identify all estimation difficulties that may occur in modeling real life data sets. In particular, we did not explore modeling with discrete covariates. Thus, the guidelines of this article should only be considered for continuous predictors. Finally, we considered only three situations—modeling with one, seven, or 25 covariates.

In conclusion, there is no single rule based on EPV that would guarantee an accurate estimation of logistic regression parameters. Instead, researchers should be careful, when fitting logistic regression models, to look at EPV as well as at the number of predictors, probable size of the regression coefficients based on previous literature, and correlations of the predictors. The simulations in this article may provide guidelines on the number of EPV necessary, given the data structure. In many situations, logistic regression modeling may pose substantial problems even if the number of EPV exceeds 10. In general, researchers should be vigilant about the correlations between their predictors and, when possible, try to increase the number of EPV.

References

[1] Maldonado G, Greenland S. Interpreting model coefficients when the true model form is unknown. Epidemiology 1993;4:310−8.

[2] Ottenbacher KJ, Ottenbacher HR, Tooth L, Ostir GV. A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. J Clin Epidemiol 2004;57:1147−52.

[3] Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

[4] Harrell F, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems and suggested solutions. Cancer Treat Rep 1985;69:1071−7.

[5] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361−87.

[6] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373−9.

[7] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007;165: 710−8.

[8] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2009.

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

[10] Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003;56:441−7.

[11] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158: 280−7.

[12] Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med 1998;17: 1623−34.