

A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions

Kenneth J. Ottenbacher^{a,*}, Heather R. Ottenbacher^b, Leigh Tooth^c, Glenn V. Ostir^d

^aDivision of Rehabilitation Sciences and Sealy Center on Aging, University of Texas Medical Branch, Galveston, TX 77555-1137, USA

^bDepartment of Nutritional Sciences, University of Arizona, Tucson, AZ, USA

^cDepartment of Social and Preventive Medicine, University of Queensland, Brisbane, QLD, Australia

^dDepartment of Internal Medicine and Sealy Center on Aging, University of Texas Medical Branch, Galveston, TX, USA

Accepted 10 May 2004

Abstract

Background and Objective: To examine if commonly recommended assumptions for multivariable logistic regression are addressed in two major epidemiological journals.

Methods: Ninety-nine articles from the *Journal of Clinical Epidemiology* and the *American Journal of Epidemiology* were surveyed for 10 criteria: six dealing with computation and four with reporting multivariable logistic regression results.

Results: Three of the 10 criteria were addressed in 50% or more of the articles. Statistical significance testing or confidence intervals were reported in all articles. Methods for selecting independent variables were described in 82%, and specific procedures used to generate the models were discussed in 65%. Fewer than 50% of the articles indicated if interactions were tested or met the recommended events per independent variable ratio of 10:1. Fewer than 20% of the articles described conformity to a linear gradient, examined collinearity, reported information on validation procedures, goodness-of-fit, discrimination statistics, or provided complete information on variable coding. There was no significant difference ($P > .05$) in the proportion of articles meeting the criteria across the two journals.

Conclusion: Articles reviewed frequently did not report commonly recommended assumptions for using multivariable logistic regression. © 2004 Elsevier Inc. All rights reserved.

Keywords: Statistical tests; Research design; Outcomes research

1. Introduction

Several studies have documented the increased use of multivariable statistical methods in the research literature examining health care [1–7]. In a review of four multivariate methods appearing in the literature from 1985 through 1989, Concato and colleagues [6] reported logistic regression was the most frequently used procedure comprising an average of 43% of the multivariate methods in the five-year period reviewed. Khan et al. [8] described a significant increase in the use of logistic regression in the obstetrics and gynecology research literature. Levy and Stolte [9] reviewed the statistical methods reported in the *American Journal of Public Health* and the *American Journal of Epidemiology*. Their survey included a probability sample of 348 articles published between 1970 and 1998. They found significant in-

creases in the use of logistic regression, proportional hazard regression, and methods for the analysis of data from complex sample surveys. Chin [10] has also reported a significant increase in the use of multivariable logistic regression in the public health and epidemiologic research literature (see Fig. 1).

Multivariable logistic regression is a sophisticated statistical procedure, and concern has been expressed regarding its application and interpretation [11–13]. The concerns have focused on assumptions associated with the appropriate use, correct interpretation, and complete reporting of multivariable logistic regression [14]. The purpose of the present study was to examine the extent to which commonly recommended assumptions and reporting requirements for multivariable logistic regression are being met in two major epidemiological journals.

2. Methods

The *Journal of Clinical Epidemiology* and the *American Journal of Epidemiology* were examined for the 2000 and

* Corresponding author. Tel.: 409-747-1637; fax: 409-747-1638.
E-mail address: kottenba@utmb.edu (K.J. Ottenbacher).

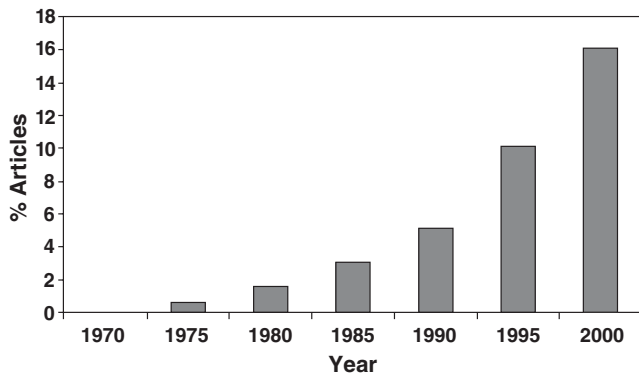


Fig. 1. Percentage of articles over 30 years in the *American Journal of Epidemiology*, *American Journal of Public Health*, *International Journal of Epidemiology*, and *Journal of Epidemiology and Community Health* with MeSH heading 'logistic regression' or with keyword 'logistic regression' or 'adjusted odds ratio.' (Data from Chin. [10])

2001 volume years. Each article was reviewed by at least two raters with experience in statistical analysis and epidemiological research to determine if multivariable logistic regression was used. All articles that reported use of multivariable logistic regression were identified. In a review of multivariate statistical methods, Concato and colleagues [6] identified five major reasons for the use of multivariable analysis. These were (a) confirmation of a relationship found in a bivariate analysis; (b) confirmation of a non-regression analysis of a relationship among multiple variables; (c) screening using large administrative databases; (d) creating a combined risk score for predicting outcomes in individual patients; and (e) quantifying risks of individual variables. Concato et al. [6] found that 73% of the multivariable analyses they reviewed (logistic regression and proportional hazard) were applied to quantify risk estimates reported as regression coefficients, odds ratios, or relative risks for individual variables. It is this application of multivariable logistic regression that we were interested in examining because, as Concato and colleagues ([6], p. 203) stated, "when individual variables are examined to determine the magnitude of their impact or risk, the estimates will vary with the structure of the mathematical model and the coding of variables. The assumptions and limitations for multivariable models then become especially important for ensuring accurate results and valid interpretation."

The reporting style in some of the articles made it difficult to determine whether the application of multivariable logistic regression involved quantifying risk estimates. In most cases, this determination was made based on the reporting and interpretation of regression coefficients, odds ratios, or relative risks for individual variables. When the two raters did not agree, a third rater reviewed the article in question and the decision agreed upon by at least two raters regarding the use of logistic regression was used in the analysis.

In spite of the high agreement between the raters (as discussed in **Results** section), the results should be viewed

as approximations of the extent to which the assumptions were met in the articles reviewed. A post hoc analysis is somewhat arbitrary in determining the nature of tests conducted. The actual appropriateness of the statistical procedures cannot be precisely determined without direct access to original data.

In identifying assumptions, we adapted criteria used by Bagley et al. [15] to evaluate multivariable logistic regression in articles examining genetic testing for cancer susceptibility. These criteria represent commonly recommended guidelines for using multivariable logistic regression [2,3]. We recognize that the assignment of recommended assumptions is somewhat arbitrary. We included assumptions that have been identified previously in the research literature as important in the reporting and interpretation of multivariable logistic regression [2,13–15]. We believe that it is important to determine if these assumptions are being followed, and if not, how that affects statistical outcomes.

Each article including multivariable logistic regression analysis was evaluated on 10 criteria associated either with the computation of the regression model and risk estimates (6 criteria), or with the reporting of the logistic regression results (4 criteria). The 10 criteria are given in Table 1, along with a brief description. Each criterion was evaluated by the two reviewers using a three-point rating scale, with 1 = criteria met, 2 = criteria not met, and 3 = cannot determine or not relevant.

A coding form similar to that used in reviewing articles for a meta-analysis [16] was developed, and included information on the basic characteristics of the article (e.g., authors, journal, year of publication, sample size) and ratings for the 10 criteria associated with the use of multivariable logistic regression (Table 1). The interrater agreement for information coded from each of the articles was examined using the intraclass correlation coefficient (ICC), model 2 as described by Shrout and Fleiss [17]. The ICC values for the 10 criteria related to the use of multivariable logistic regression ranged from 0.91 to 0.99.

3. Results

Fifty-one articles in volume 54 (2001) of the *Journal of Clinical Epidemiology* and 45 articles in volumes 151/152 (2000) and 153/154 (2001) of the *American Journal of Epidemiology* met the inclusion criteria. The final sample included 99 articles. A complete list of the articles is available from the first author.

The percentage of articles meeting each of the 10 criteria for the two journals is presented in Fig. 2. There were no significant differences between the two journals in the percent of articles meeting any given criterion. The information from the two journals is therefore combined in the descriptions that follow.

Table 1

Criteria used to examine logistic regression models in articles reviewed in the *Journal of Clinical Epidemiology* and *American Journal of Epidemiology*

Criterion	Description
Sufficient events per independent variable	The ratio of outcome events to independent variables should be 10:1 or higher. The fewer events per independent variable, the greater the opportunity for the estimates of the regression coefficients to be unreliable; the sample variance of the model coefficients, and confidence intervals will also be less accurate.
Conformity with linear gradient for continuous variables	Articles with continuous or ranked independent variables test to assure conformity with the linear gradient or check on the log-odds scale. This is not an issue for dichotomous predictor variables for which there are only two values and one possible change.
Tests for interactions	Article includes a discussion of interaction terms and why they were either included or not included. If interactions are included, then the significance of the interaction is reported.
Collinearity	Explicit tests for collinearity are undertaken and reported. Some software packages may include automatic checks for collinearity – if so the fact the collinearity was examined is reported.
Validation	Model validation is discussed and validation procedure reported if appropriate, e.g., split-sample methods, cross validation, bootstrapping or other resampling methods.
Statistical significance	Statistical tests of significance are applied to each variable's coefficients and to the entire model.
Goodness-of-fit, Discrimination measures	Summary goodness-of-fit measures or discrimination statistics (ROC curves) are reported describing how well the entire model matches the observed values.
Selection of independent variables	Article explains how variables were selected for inclusion into the model? Variables may be chosen based on earlier research; sometimes they are selected by virtue of significant association in a bivariate analysis with the outcome variable.
Coding of variables	Study provides an appropriate description of the coding for independent variables. The coefficient for an independent variable depends on how that variable is coded. The effect of the coding on the interpretation of the regression coefficients is especially important when interaction terms are reported.
Fitting procedure	Procedure for entering variables into the model is explicitly stated, with description of appropriateness of method selected (e.g., forward inclusion, backward elimination, best-subset, or specified a priori, either collectively or in "hierarchically" grouped subsets).

3.1. Sufficient events per variable

The criterion was set at a ratio of 10:1 (events per independent variable). Simulation studies have indicated that regression coefficients and other indicators of risk may be unreliable if fewer than 10 events per independent variable are included in the model [18,19]. The ratio of events to independent variable in the studies ranged from 2.6 to 26.2. Forty (40%) of the 99 articles met the criterion of a 10:1 ratio of events to independent variable for the least common outcome in the study.

3.2. Linear gradient of ranked and continuous variables

The majority, 63 (63%) of the 99 articles used only binary independent variables in the logistic regression models and the criterion for nonconformity to linear gradient did not apply. Some binary variables as reported in the reviewed articles may have represented continuous or ordinal variables that were dichotomized. The actual pattern of binary versus continuous variables cannot be determined without access to the original data. In the 36 (36%) articles reporting logistic regression models with ranked or continuous independent variables, 29 (81%) provided no information or indication of checking for nonconformity to a linear gradient.

3.3. Tests for interactions

Thirty-nine (39%) studies reported or discussed testing for interactions among the independent variables. In the

remaining investigations, it was not clear whether interactions were tested and not found to be significant and therefore not reported, or whether no tests for interactions were conducted.

3.4. Collinearity

Seventeen (17%) of the studies provided some discussion or description of testing for multicollinearity among the independent variables. In 12 cases, this description involved a brief statement that the statistical package used to analyze the data computed a covariance matrix and included procedures to test for collinearity among variables.

3.5. Validation

Only three (3%) of the 99 studies reported validation procedures for predictive models using a subset of the original sample (50–60%) to develop a model and coefficients, and then tested the results against the remaining (40–50%) subjects and responses. Six (6%) studies reported some form of bootstrapping or resampling method in developing the models.

3.6. Statistical significance

Statistical significance (*P*-values), confidence intervals, or both were reported in 100% of the articles. *P*-values were reported in 84% of the articles and confidence intervals in 71%.

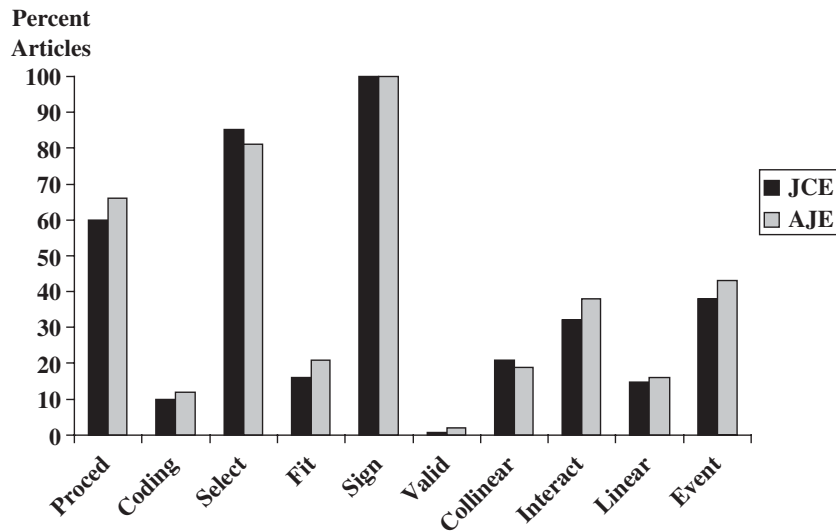


Fig. 2. Comparison of 10 criteria for using multivariable logistic regression as reported in the *Journal of Clinical Epidemiology* (JCE) and the *American Journal of Epidemiology* (AJE) for 2000 and 2001. Figure shows percent of articles meeting each criterion. See Table 1 for explanation of the 10 criteria (y-axis) rated for each journal.

3.7. Goodness-of-fit and discrimination

Goodness-of-fit statistics, discrimination statistics (ROC curves), or regression diagnostics were reported in 19 (19%) of the articles. The most commonly reported measure was the Lemeshow–Hosmer goodness-of-fit summary statistic. The goodness-of-fit statistic was most frequently applied to the overall model. In only 6 (6%) of the articles were goodness-of-fit statistics reported for independent variables.

3.8. Selection of independent variables

Eighty-one (81%) of the 99 articles were rated as providing an adequate description of the rationale and procedures for selecting independent variables included in the multivariable logistic regression models.

3.9. Coding of independent variables

Ten articles (10%) were judged as providing a complete description of the coding for all independent (predictor) variables included in the multivariable logistic regression models.

3.10. Model fitting procedure

Sixty-five (65%) of the studies reported adequate information on the model fitting procedures used. In these articles, the procedure for entering variables into the model was stated with a description of the rationale.

4. Discussion and conclusions

As the use of multivariate procedures continues to increase in the health and epidemiological literature, it is important that the procedures be applied correctly and reported

completely [20,21]. We examined the recent application and reporting of multivariable logistic regression in two major epidemiological journals and found several areas where the application and description could be improved. Our results confirm and extend those reported previously regarding the use of multivariate statistical methods (including multivariable logistic regression) in the medical and health care literature [10,11,15]. In particular, we found that only 3 of the 10 criteria were addressed in at least 50% of the reviewed articles. The three criteria were (a) testing for statistical significance or providing confidence limits for independent (predictor) variables, or both (100%); (b) describing rationale and methods for selecting independent (predictor) variables included in the model (82%); and (c) describing the procedure or procedures used to generate the model (65%). Criteria for which fewer than 20% of the reviewed articles were considered adequate included (a) describing conformity to a linear gradient (19%); (b) examining or reporting on collinearity (17%); (c) reporting information on validation procedures (3–6%) or goodness-of-fit or discrimination statistics (19%); and (d) providing complete information on coding the independent variables (10%). Other areas of concern where fewer than 50% of the articles met the criteria included reporting or discussing interaction terms (39%) and meeting the events per independent variable ratio threshold of 10:1 (40%).

Not all assumptions will be viewed as equally important, and there is disagreement among statistical experts regarding how some assumptions and requirements should be operationalized [22]. For example, various indices have been proposed for assessing the goodness-of-fit, reflecting how well the logistic regression model fits the actual data [23]. There are also differing points of view on methods to test for collinearity, the validity of the model, and nonconformity

to a linear gradient [24,25]. There is general agreement regarding the importance of validation procedures for developing predictive models. The efficiency of external versus internal validation methods, however, is open to debate. With moderate sample sizes, the precision of risk estimates and power of hypothesis testing will be reduced by using a subset of the data. This may, partially, explain the small number of studies that reported using this validation procedure. Some of the concerns related to sample size can be addressed by the use of bootstrapping or other resampling techniques; however, Bleeker and colleagues [26] recently demonstrated the limitations of bootstrapping as a method of validation for prediction models with smaller data sets.

Although there is some question about what the appropriate ratio of events to variables should be in predictive models versus other applications, there is general agreement that this is an important assumption that, if not adequately addressed, will have a negative effect on the statistical results [18,19]. The failure to meet this assumption is often discussed in terms of regression models that are “overfit” or “underfit” [27]. The usual tests of statistical significance may be invalid and confidence intervals associated with individual risk estimates difficult to interpret [28] in models with too few events per variable (overfit). Logistic regression models may also be underfit, when the power to detect important relationships is low due to a scarcity of outcome events and essential variables are omitted from the model. The problem of model overfit or underfit has been reported in the epidemiological and statistical literature [1,2,14], but the implications are not widely recognized or well understood. For example, most “instructions for authors” do not identify statistical requirements for reporting multivariate statistics. One exception is the *Manual of Style* for the American Medical Association, which does include a requirement that the steps used to develop a multivariate model be reported, and that a sample size of 25 individuals or greater be included for each independent variable in multivariable logistic regression [29]. The key issue, however, is not the sample size, per se; rather, it is the appropriate number of outcome events for each independent variable.

Although our results are consistent with previous findings regarding the use and reporting of multivariate procedures in related areas [6,8,14], our investigation has limitations. For example, the results are not generalizable to other journals. We examined a relatively small sample of recent articles appearing in two respected journals. These findings do not reflect an examination of multivariable logistic regression as used or reported in the broader health care or epidemiologic literature. Nonetheless, we have no reason to believe that the sample of studies we examined is unique or not representative of the epidemiological research literature.

Another limitation of our investigation, previously noted, is the selection of statistical assumptions and reporting issues. We attempted to identify assumptions and concerns that have been reported in the statistical and epidemiological literature [6,8,9,11,14,15], and we operationalized these as

the 10 criteria examined in the present study. There will be disagreements with the criteria we selected (or failed to select) and with how they were operationalized and rated. For example, we did not include a criterion related to the effect of outliers or influential observations on the results of multivariable logistic regression. A small number of observations can have a profound influence on the results of logistic regression if the observations are distinctly different from all others [1]. Although mathematical models exist for dealing with these influential observations (outliers), it is often difficult to determine if such observations are statistical artifacts or reflect an important biological or other relation with the independent variable [1,2]. Other investigators may disagree with the way a criterion was rated. In evaluating the criterion on Fitting Procedures (Table 1), we were interested in whether the authors adequately described a procedure for entering variables into the model. We realize that some methods of entry will be seen as better than others, and this may depend on the question being studied. We did not attempt to make judgments about the appropriateness of the entry method (e.g., backward, forward, stepwise).

The validation criterion used (Table 1) might not apply to all studies included in the review. In some cases, it was difficult to determine if prediction was a goal of the investigators. The assumption that validation was relevant to all studies was conservative and may have produced an artificially low percent of studies not meeting this criterion. We are confident that the majority of the reviewed studies included a predication component. If the denominator is changed from 99 to 50, the percent of studies meeting this criterion would remain low, at 10–12%. We are also aware that significance testing approaches for selecting and entering variables may yield invalid point and interval estimates [30]. Our goal was to provide information on whether the criterion of describing the method for variable entry was met, not on judging the appropriateness of the method for a particular research question.

Finally, as noted in a previous section, any analysis of assumptions involving published reports is limited by not having access to original data. For some of the criteria included in this analysis, it was not possible to determine if the criteria were truly not met, or if the authors simply failed to report the results or document their procedures. For example, some investigators may have tested for interactions among independent variables included in their models, and having found none, not reported on the tests for interactions.

Despite these limitations, we believe the results suggest areas that can be improved in the teaching, application, and reporting of multivariable logistic regression. These include more complete reporting, as in our example dealing with testing for interactions, or providing detailed information regarding the coding of the independent variables included in the model. Goodness-of-fit and discrimination statistics are computed in most statistical software packages and providing these values would assist readers in interpreting the results from logistic regression models. Conforming to the

events-to-independent variable ratio of 10:1 would require investigators to be more circumspect in the variables included in logistic models, or explain how not meeting this criterion may have influenced the results.

There is a continuing need in health and epidemiological research to formulate concise questions and describe the results of quantitative analyses completely and accurately. Researchers must continue to interpret and report the results of their investigations as accurately and completely as possible. Knowledge of statistical assumptions and improved reporting requirements can help achieve this goal.

Acknowledgments

This research was supported by a grant from the National Institutes of Health, U.S. Department of Health and Human Services, Independent Scientist Award (KO2-AG019736).

References

- [1] Hosmer DW, Lemeshow S. Applied logistic regression. 2nd edition. New York: Wiley; 2000.
- [2] Feinstein AR. Multivariable analysis: an introduction. New Haven, CT: Yale University Press; 1996.
- [3] Corbie-Smith G, Viscoli CM, Kernan WN, Brass LM, Sarrel P, Horwitz RI. Influence of race, clinical, and other socio-demographic features on trial participation. *J Clin Epidemiol* 2003;56:304–9.
- [4] Stel VS, Smit JH, Pluijm SMF, Lips P. Balance and mobility performance as treatable risk factors for recurrent falling in older persons. *J Clin Epidemiol* 2003;56:659–68.
- [5] Vollmer RT. Multivariate statistical analysis for pathologists. Part I, the logistic model. *Am J Clin Pathol* 1996;105:115–26.
- [6] Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 1993;118:201–10.
- [7] Katz MH, Hauck WW. Proportional hazards (Cox) regression. *J Gen Intern Med* 1993;8:702–11.
- [8] Khan KS, Chien PF, Dwarakanath LS. Logistic regression models in obstetrics and gynecology literature. *Obstet Gynecol* 1999;93:10014–20.
- [9] Levy PS, Stolte K. Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. *Stat Methods Med Res* 2000;9:41–55.
- [10] Chin S. The rise and fall of logistic regression. *Aust Epidemiol* 2001; 8:7–10.
- [11] Hall GH, Round AP. Logistic regression: explanation and use. *J R Coll Physicians Lond* 1994;28:242–6.
- [12] Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996;49:907–16.
- [13] Bender R, Grouven U. Logistic regression models used in medical research are poorly presented [Letter]. *BMJ* 1996;313:628.
- [14] Campillo C. Standardizing criteria for logistic regression models. *Ann Intern Med* 1993;119:540–1.
- [15] Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical condition. *J Clin Epidemiol* 2001;54:979–85.
- [16] Cooper HM. Synthesizing research: a guide for literature reviews. 3rd edition. Thousand Oaks, CA: Sage; 1998.
- [17] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing reliability. *Psychol Bull* 1979;86:420–8.
- [18] Concato J, Feinstein AR. Monte Carlo methods in clinical research: applications in multivariable analysis. *J Investig Med* 1997;45:394–400.
- [19] 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.
- [20] Hirsch RP, Riegelman RK. Statistical first aid: interpretation of health research data. Boston: Blackwell Scientific Publications; 1992.
- [21] Lang TA, Secic M. How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers. Philadelphia, PA: American College of Physicians; 1997.
- [22] Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treat Rep* 1985;69:1071–7.
- [23] Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Public Health* 1991;81:1630–5.
- [24] Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med* 1991;10:1213–26.
- [25] Begg MD, Lagakos S. On the consequences of model misspecification in logistic regression. *Environ Health Perspect* 1990;87:69–75.
- [26] Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, Moons KGM. External validation is necessary in prediction samples: a clinical example. *J Clin Epidemiol* 2003;56: 826–32.
- [27] Katz MH. Multivariable analysis: a practical guide for clinicians. New York: Cambridge University Press; 1999.
- [28] Lemeshow S, Hosmer DW. Logistic regression. In: Armitage P, Colton T, editors. *Encyclopedia of biostatistics*. New York: Wiley; 1998:2316–27.
- [29] Iverson C; American Medical Association. *American Medical Association manual of style: a guide for authors and editors*. 9th edition. Baltimore, MD: Williams & Wilkins; 1998.
- [30] Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic analysis. *J Clin Epidemiol* 1999;52:935–42.