Medicine®

OBSERVATIONAL STUDY

OPEN

Quality Reporting of Multivariable Regression Models in Observational Studies

Review of a Representative Sample of Articles Published in Biomedical Journals

Jordi Real, BSc, Carles Forné, MSc, Albert Roso-Llorach, MSc, and Jose M. Martínez-Sánchez, PhD

Abstract: Controlling for confounders is a crucial step in analytical observational studies, and multivariable models are widely used as statistical adjustment techniques. However, the validation of the assumptions of the multivariable regression models (MRMs) should be made clear in scientific reporting. The objective of this study is to review the quality of statistical reporting of the most commonly used MRMs (logistic, linear, and Cox regression) that were applied in analytical observational studies published between 2003 and 2014 by journals indexed in MEDLINE.

Review of a representative sample of articles indexed in MEDLINE (n = 428) with observational design and use of MRMs (logistic, linear, and Cox regression). We assessed the quality of reporting about: model assumptions and goodness-of-fit, interactions, sensitivity analysis, crude and adjusted effect estimate, and specification of more than 1 adjusted model.

The tests of underlying assumptions or goodness-of-fit of the MRMs used were described in 26.2% (95% CI: 22.0–30.3) of the articles and 18.5% (95% CI: 14.8–22.1) reported the interaction analysis. Reporting of all items assessed was higher in articles published in journals with a higher impact factor.

A low percentage of articles indexed in MEDLINE that used multivariable techniques provided information demonstrating rigorous application of the model selected as an adjustment method. Given the importance of these methods to the final results and conclusions of observational studies, greater rigor is required in reporting the use of MRMs in the scientific literature.

(*Medicine* 95(20):e3653)

Abbreviations: CI = confidence interval, MRM = multivariable regression model.

Editor: Zhiyong Liu.
Received: October 8, 2015; revised: April 7, 2016; accepted: April 18, 2016.
From the Unitat de Suport a la Recerca-Lleida, Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Barcelona (JR); Universitat International de Catalunya, Facultat de Medicina i Ciències de la Salut, Sant Cugat (JR, JMM-S); Department of Basic Medical Sciences, Universitat de Lleida, Lleida (CF); Oblikue Consulting (CF); Institut Universitari d'Investigació en Atenció Primària Jordi Gol, Barcelona (AR-L); and Tobacco Control Unit, Catalan Istitute of Oncology, Hospitalet de Llobregat (JMM-S), Spain.
Correspondence: Dr. Jose M. Martínez-Sánchez, Departament de Ciències Bàsiques, Universitat Internacional de Catalunya, Carrer de Josep Trueta s/n, 08195 Sant Cugat del Vallès, Barcelona, Spain (e-mail: jmmartinez@uic.es).
The authors have no conflicts of interest to disclose.
Supplemental Digital Content is available for this article.

ISSN: 0025-7974
DOI: 10.1097/MD.0000000000003653

INTRODUCTION

Two important aspects of biomedical research are the internal and external validity of the study design.[1] Information bias and confounding variables affect internal validity and are present to some extent in all observational research. Information bias results from incorrect determination of the exposure, outcome, or both. Confounding is a ''mixture'' or ''diffusion'' of effects: a researcher attempts to associate an exposure with a result, but actually measures the effect of a third—sometimes unnoticed—factor, that is, a confounding variable. This bias can be diminished, but only if the confounding factor is anticipated and the relevant data are collected to allow proper adjustment.

Confounding factors can be controlled in various ways: restriction, matching, stratification, standardization, and multivariable techniques. All of these approaches are focused on achieving homogeneity between study groups,[1] and in recent years multivariable regression models (MRMs) such as linear, logistic, Poisson, or Cox regression have become popular and very frequently used.[2] A review of research based on Canada National Health Survey data found that nearly 80% of the studies used some type of MRM, predominantly logistical modeling.[3] A systematic review of studies published by 10 prestigious journals in epidemiology and general medicine showed that almost 95% used MRMs, in addition to other techniques, as the adjustment methodology.[4] The frequency with which any statistical method is applied is often determined by the available software and computational capacity; therefore, this high rate of MRM usage could be due to major advances in computational capabilities with increased availability of data, but also to the ease with which these techniques can now be applied using standard statistical software.[5]

An advantage of MRM analysis is that it allows the control of more confounding factors, compared to stratification, and a simultaneous evaluation of the relationship between several exposure factors and response variables of different types (continuous, dichotomous, count, or time-dependent events).[2,6] The estimated effect of each variable reflects its association with the outcome, taking into account the contribution of the rest of the variables introduced into the model. However, modification effect is not identifiable by simple inclusion of the variable in the regression model; the interaction terms between exposure and effect modification, or the confounding variable, must also be included.

Moreover, MRMs assume probability distributions that include underlying assumptions (e.g., assumptions of normality, homoscedasticity, independence of errors, etc.). In addition, parameter estimation could be inefficient if there is multicollinearity between 2 or more variables, which affects convergence in the inference process, among other potential problems.[5,6]

Regression models produce nonbiased results for each variable of interest if the model is correctly specified and all potential confounding factors are included and correctly measured.[7] Furthermore, if not all confounders are included or the model is not properly specified, the consequences are residual confounding and biased estimates.[8,9] Although the underlying ''true'' model is seldom known, specification errors and residual confounding can be minimized by testing the formal assumptions of the selected model.[6,10] Specific statistical tools are available to evaluate whether all necessary conditions have been met to apply a particular type of adjusted modeling and the appropriateness of the model that was finally selected.[6,10] In addition, given that MRMs are usually sensitive to model specification, it is desirable to carry out more than 1 adjustment strategy to evaluate the stability of the estimated effects of different settings.[11] All these measures, together with a sensitivity analysis (variation by subgroups) and interaction assessment, lead to more consistency in evaluating the adjusted measures of association and increase their validity and level of evidence.[12,13]

Various studies have described the statistical methodology used in published biomedical research.[3,4,14,15] Strasak et al[15] showed that inappropriate use of some statistical tests is one of the most common errors. In 2008, Groenwold et al[4] carried out a systematic review of observational studies published in general medical and epidemiology journals with a high impact factor and reported finding poor quality in the adjustment methods used. More recently, in 2014, another systematic review of the use and application of generalized linear mixed models showed their increased use and, at the same time, room for improvement in reporting quality.[14] However, there is a lack of evidence on the quality of reporting or the validation procedures used when MRMs are applied in observational studies. Therefore, the objective of the present study was to review the quality of statistical reporting when the most commonly used MRMs (logistic, linear, and Cox regression) were applied in analytical observational studies published between 2003 and 2014 by journals indexed in MEDLINE.

MATERIALS AND METHODS

We reviewed a representative random sample of articles indexed by MEDLINE using the PubMed search engine. The search was specifically designed to identify original studies with an analytical observational design that stated their use of logistic, Cox, or linear MRMs focused on confirmatory analysis (i.e., to assess the effect of exposure) (Supplementary Table S1, http://links.lww.com/MD/A979 of the Appendix). The search was limited to studies in humans that were published in English between January 1, 2003 and February 16, 2014. Clinical trials, editorials, commentaries, and case reports were excluded. This strategy retrieved 71,519 references, from which a simple random sample of 500 articles was selected. A sample size of 500 randomly selected papers was calculated to allow estimation with 95% confidence and a precision of ±5% units, a population percentage considered to be of 50%. We assumed a 50% prevalence to maximize the sample size. A replacement rate of 20% was anticipated. Exclusion criteria removed 72 references, including those that proved to be focused on diagnosis, prognosis, or other analytical approaches. Therefore, 428 papers were finally reviewed (Figure 1).

Items Reviewed in Full-Text Analysis

Based on the literature,[2,11,16,17] a list of aspects related to the application of MRMs was specified, including testing formal assumptions, goodness of fit: interactions, and sensitivity analysis of the adjustment models (Table 1). An initial review of 10 articles served as a pilot test for the entire research team to define the list of items to be included, establish precise definitions, and improve interrater homogeneity. Finally, the definitive set of MRM-related items to be verified in each relevant section of the manuscript was established (Table 1). Each item was classified according to whether it would likely be stated in the methods section or was mainly involved in the communication of findings and would appear in the results section. If an item was reported in any section, the paper was considered to meet the criteria.

Review Procedure

The selected articles were randomly distributed among the 3 designated reviewers on the research team. Any doubts were shared and resolved by consensus. In addition, 12 articles were randomly selected for reviewing by all 3 reviewers, for blinded evaluation of interrater agreement. No significant differences in outcomes between reviewers were observed (Supplementary Table S2, http://links.lww.com/MD/A979); there was high interrater agreement (Kappa index > 0.73) and intraclass correlation coefficient of the number of completed items (0.88). The Kappa index measures the agreement between reviewers for compliance with each of the items separately (dummy variables) and the intraclass correlation coefficient quantifies the correlation of the number of completed items (numerical variable) between reviewers. A detailed analysis of intra- and interrater agreement is shown in the Appendix, Supplementary Table S2, http://links.lww.com/MD/A979.

Statistical Analysis

For each item specified for review, prevalence estimates and 95% confidence interval (CI) were obtained. We also calculated mean and standard deviation (SD) for the total number of review items fulfilled. CIs were computed using normal approximation. All analyses were stratified in groups according to the impact factor of the journal in the year of publication ($\leq$2, 2–4, $>$4), sample size ($<$500, 500–1500, $\geq$1500), design (cross-sectional, cohort, and case–control), data source (ad hoc, clinical/administrative records, both, or ''mixed''), and type of MRM (logistic, linear, and Cox). Pearson χ^2 and trend tests were used to assess the association between prevalence of the items of interest and the categorical secondary variables. Mann–Whitney *U* test was used to examine the relationship between prevalence of items, sample size, and the journal's impact factor. We computed the 2-sided criteria for all variables and 1-sided criteria for the impact factor level because the higher the impact factor, greater rigor is required in reporting the use of MRMs in the scientific literature. To assess and control for possible interactions, the analysis was again stratified by impact factor, sample size, design, and type of modeling. Significance level was set at $\alpha = 0.05$. All analysis was carried out using the SPSS statistical software, version 18.0. (PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc.).

Ethical statement: None required the approval of the Ethics Committee because the primary source was secondary data from published scientific articles.

RESULTS

Of 428 articles reviewed, published in 313 journals (mean of impact factor = 3.38), 49.5% were cohort studies, with data primarily collected using questionnaires specifically designed

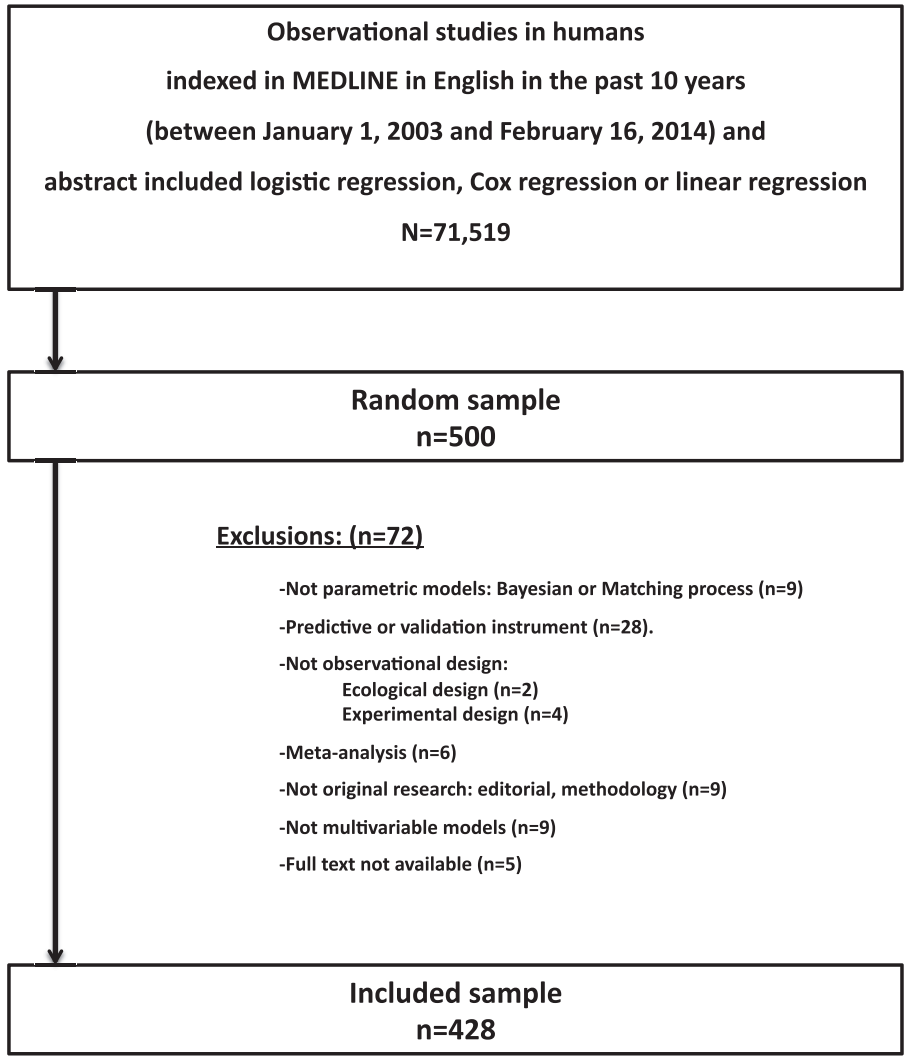

FIGURE 1. Flowchart of articles included.

to address the research objective (45.6%). The most frequently used type of modeling was logistic regression (67.5%), followed by Cox (22.9%) and linear (18%) regression. Nearly half (48.8%) of the articles reviewed were published during the last 3 years of our study period (2010–2013). Only 4% of the articles referenced any other publication that expanded on the methodology used in the study.

Table 2 shows the overall percentage observed for each of the items reviewed, and in relation to all selected variables. The major item that was reported most often (33.4%; 95% CI: 28.9–37.8%) was ''crude and adjusted effect'' (item 4, Table 2), followed by sensitivity analysis (32.7%; 95% CI: 28.3–37.1%). The least-reported item was interaction analysis (18.5%, 95% CI: 14.8%–22.1%). Testing the assumptions of the model and fitting more than 1 model were reported in 26.2% and 25.7% of the articles, respectively (items 1 and 5, Table 2).

The percentages observed for all of the items analyzed were higher in studies published in journals with a moderate or high impact factor (Table 2). The assessment of model adjustment criteria (item 1) was primarily observed in articles published in journals with a moderate impact factor and in studies that used linear models. The criteria referring to interactions, sensitivity analysis, and testing more than 1 model (items 2, 3, and 5, respectively) were also significantly and directly associated with sample size (Table 2).

The mean number of items identified in the articles reviewed was 1.36 (SD = 1.17), and increased with sample size and impact factor ($P < 0.001$). Both factors act independently of the mean number of items: there was no observed interaction between impact factor and sample size. Figure 2 shows how the frequency of each item increased with impact factor (Figure 2A), independently of sample size (Figure 2B), study design (Figure 2C), and type of MRM used (Figure 2D).

DISCUSSION AND CONCLUSIONS

Our study shows very low reporting of MRM validation in observational studies indexed in MEDLINE, being higher in studies with larger sample sizes published in journals with a higher impact factor. Only 26.2% of the articles reviewed described their validation analysis of assumptions or goodness-of-fit for the MRM used, 33.4% showed both the crude

TABLE 1. Primary Items Reviewed in Manuscripts of Observational Studies That Used Multivariable Methods (Logistic Regression, Cox Regression, or Linear Regression)

Item	Issues Reviewed in the Manuscript (Yes/No) Detail and Justify the Item Reviewed	Method Section	Results Section
1	Model assumptions and goodness-of-fit*: Normality, linearity/log-linearity, homoscedasticity, proportional hazards assumption (Cox models) or goodness-of-fit. Is the functional form of the selected model correct? How far away from the data is the selected model?	x	X
2	Interaction analysis: Some interaction term was evaluated in the models. Is there any potential variable that can modify the estimated effect?	x	x
3	Sensitivity analysis: Sensitivity analysis of the models was performed with subsamples Are the findings sufficiently robust, considering the process used to obtain them?	x	x
4	Crude and adjusted effect estimate: Report of crude measures of association in addition to those adjusted according to the model used (Odds ratio, hazards ratio, etc.). How much does the studied effect change when other variables are taken into account?		x
5	More than one adjusted model specified: For each response variable, more than 1 adjusted model with different combinations of variables was shown. Does the estimated effect differ between the different adjusted models, settings, specifications, etc.?		x

*Specific statistical methods: Kolmogorov–Smirnov about residuals, Q–Q plots; Hosmer–Lemeshow test for logistic regression, Schoenfeld residuals for Cox Regression, R^2, receiver operating characteristic (ROC) curve.

and adjusted effects, and 32.7% described any sensitivity analysis. Interaction analysis was only observed in 18.5% of the articles reviewed.

Our results are consistent with previous scientific evidence.[4,14,18] Müllner et al[18] showed that journals with a higher impact factor had better statistical reporting, perhaps because their editorial process specifically includes statistical review. In our study, the percentages observed for all of the items analyzed were higher in studies published in journals with a higher impact factor. A systematic review by Casals et al[14] of 108 articles that applied generalized linear mixed models, without discriminating between type of design or research objective, found that validation of the model and testing for goodness-of-fit were reported in 6.5% and 15.7%, respectively, of the articles. In contrast, our results showed a higher prevalence of this item (17.7%–33.7%, depending on the impact factor of the journal). This difference could be explained because our review is based on a random sample that included methodologies whose use is much more widespread.[19] Another systematic review found a lack of attention to adjustment methods in analytical observational studies,[4] in contrast with diagnostic, prognostic, or predictive validation studies in which combinations of variables were modeled with greater precision.[16,20–23] In the latter types of studies, calibration, discriminatory power, goodness-of-fit, and validation of the statistical model are considered essential 1st steps before selecting the final adjusted model.[16] The recent Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement[24] provides guidelines that highlight the essential aspects of developing and validating a predictive multivariable model. Requirements of this guide include, among others, the need for using internal validation methods to evaluate model's performance and to compare multiple models.[24]

On the other hand, published guidelines provide specific recommendations on the reporting of the scientific results of clinical trials (CONSORT),[25] observational studies (STROBE),[26] or statistical analysis in general scientific literature (SAMPL).[27] These guidelines were developed to provide more complete and precise information about key aspects of research studies, and some have been incorporated into the author guidelines of major scientific journals. Nonetheless, even though the STROBE guidelines highlight the control of confounders as a crucial aspect of observational studies and the SAMPL and TRIPOD guidelines broaden the standards for the scrutiny of statistical methods, there is still a void in requiring or assessing multivariable methodology in observational designs. Notably, the Guide for Authors and Editors (Manual of Style for the American Medical Association) includes the need to report model diagnostics and proportion of variance explained by both individual variables and the complete model.[28] In this sense, even though the data analysis may be correct, inadequate reporting makes it impossible for the reader to assess whether the data were processed appropriately.[18]

In observational research, best practice includes avoiding bias in the study design, adjusting for possible bias in the data analysis if it is not possible to avoid bias entirely in the design, and quantifying and analyzing the effects of residual bias on the study results.[7] Nonetheless, if the model was not properly selected, there may be major residual confounding even after MRM adjustment,[29–31] which leads to bias in the associations studied. For example, Liang et al[8] recently published the results of a simulation study, concluding that ''even when all confounding factors are known and controlled for using conventional multivariable analysis, the observed association between exposure and outcome can still be dominated by residual confounding effects.''

TABLE 2. Frequency of Items Related to the Application of Statistical Models, Based on Study Characteristics in Articles Reviewed

			Percentage					Reporting at Least Item
Variable	**Category**	**N**	**it1**	**it2**	**it3**	**it4**	**it5**	**≥1**
Overall		428	26.2%	18.5%	32.7%	33.4%	25.7%	71.5%
	95% CI		22.0–30.3	14.8–22.1	28.3–37.1	28.9–37.8	21.6–29.8	67.2–75.8
Design*								
	Cross-sectional	108	31.5%	12.0%	33.3%	30.6%	17.6%	68.5%
	Cohort	212	26.9%	20.8%	32.1%	36.3%	26.4%	72.6%
	Case–control	72	16.7%	25.0%	43.1%	23.6%	37.5%	75.0%
	P-value†		0.083	0.065	0.229	0.123	0.011	0.603
Data source*								
	Ad hoc	195	27.2%	16.4%	31.8%	33.8%	25.1%	69.2%
	Clinical record	106	24.5%	16.0%	30.2%	32.1%	20.8%	69.8%
	Mixed	114	25.4%	26.3%	37.7%	34.2%	31.6%	78.9%
	P-value†		0.870	0.067	0.437	0.936	0.179	0.155
Sample size								
	<=500	205	26.8%	12.2%	21.5%	30.2%	18.5%	62.0%
	501–1500	90	25.6%	21.1%	35.6%	43.3%	31.1%	75.6%
	1501+	133	25.6%	26.3%	48.1%	31.6%	33.1%	83.5%
	P-value†		0.956	0.004	<0.001	0.078	0.005	<0.001
	P-value‡		0.786	0.001	<0.001	0.634	0.002	<0.001
	P-value§		0.780	<0.001	<0.001	0.543	0.002	<0.001
Impact factor of journal								
	<=2.00	147	17.7%	8.8%	17.7%	29.3%	17.7%	55.1%
	2.01–4.00	166	33.7%	18.1%	34.3%	34.3%	25.9%	76.5%
	4.01+	115	26.1%	31.3%	49.6%	37.4%	35.7%	85.2%
	P-value†,‖		0.003	<0.001	<0.001	0.185	0.002	<0.001
	P-value‡,‖		0.040	<0.001	<0.001	0.0.08	<0.001	<0.001
	P-value§,‖		0.033	<0.001	0.001	0.229	<0.001	<0.001
Model¶								
	Logistic	289	14.9%	14.2%	26.6%	28.7%	22.1%	67.8%
	Linear	77	39.0%	14.3%	32.5%	22.1%	18.2%	87.0%
	Cox	98	29.6%	19.4%	22.4%	32.7%	20.4%	72.4%
	P-value†		<0.001	0.343	0.081	0.462	0.987	0.006

Description of items 1–5: it1 = model assumptions and fit, it2 = interaction analysis, it3 = sensitivity analysis, it4 = crude and adjusted effect estimates, it5 = more than one adjusted model. CI = confidence interval.

*The category with undefined information is not included (36 design manuscripts and 13 of data source).

†*P*-value computed with χ^2 Pearson test.

‡*P*-value computed with χ^2-trend test.

§*P*-value computed with Mann–Whitney *U* test.

‖*P*-value computed with unilateral test.

¶Statistical test excluded 35 manuscripts with multiple models.

In this sense, overall goodness-of-fit, along with graphical validation analysis, can allow researchers to evaluate possible conflicts between the models and alert them to possible specification problems, even without ensuring that the model is completely correct.[32] Properly fitting the model may require additional adjustment variables, their transformation, inclusion of interactions, or the choice of other adjustment techniques that are less sensitive to the selection of a particular model, such as stratified analysis, matching techniques,[33] or other flexible modeling approaches.[34]

Strengths and Limitations

One of the limitations of this analysis is that the primary source of data was the abstracts accessed in MEDLINE by the PubMed search engine. Therefore, the universe of potential studies for analysis was limited to that repository and the sensitivity of the search strategy used. In an effort to minimize missed records, we designed a highly specific search-term strategy. During our review we only had to discard 14.4% of the manuscripts for failing to meet at least 1 inclusion criterion.

Another limitation was that the quality or transparency of the methodological reporting could be affected by the word limitations imposed by a journal's guidelines. Nonetheless, it is now usually possible to complement an article with online supplementary information or to disseminate the methodological details and protocols in a separate manuscript that provides greater detail about the more technical aspects. However, only 4% of the manuscripts included in the present review contained any reference to a separate article detailing the methodology used.

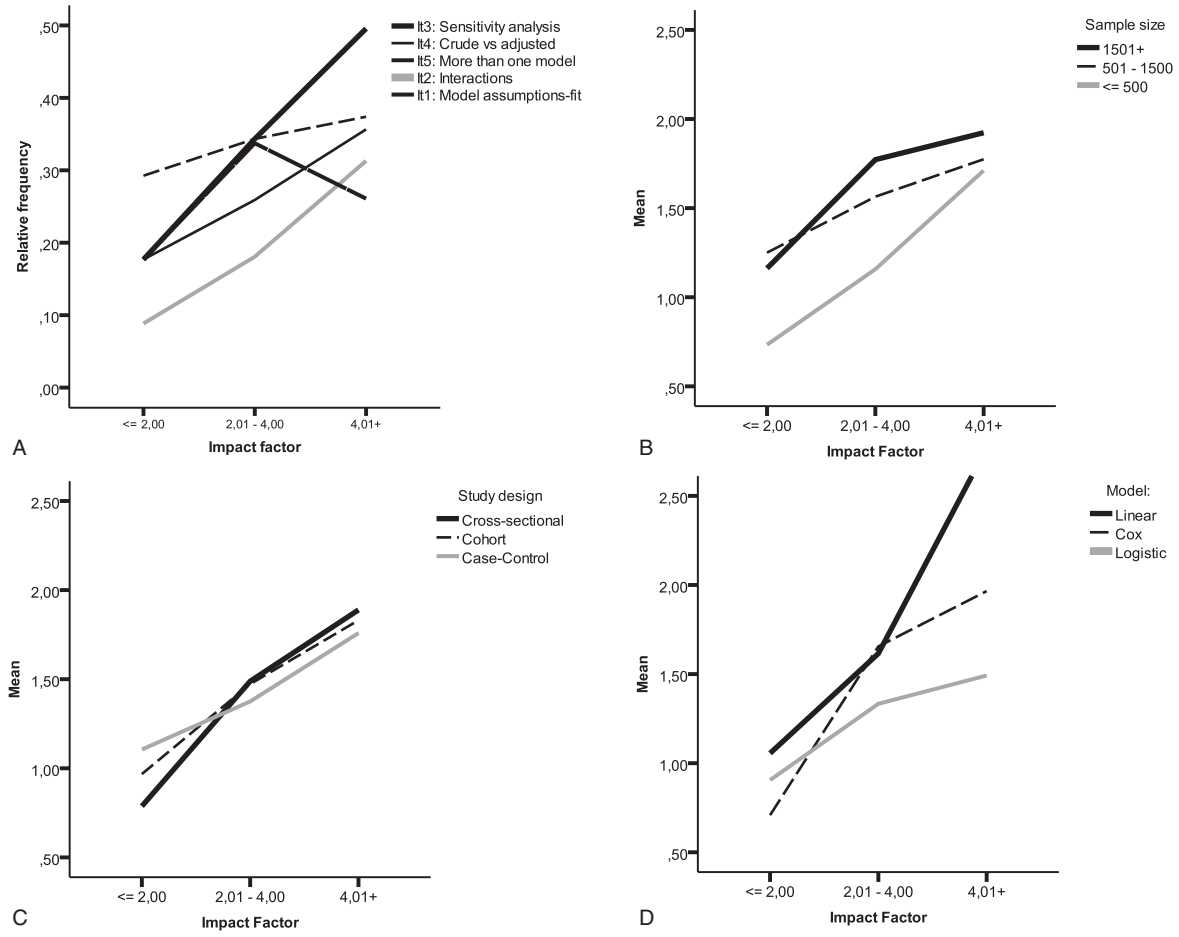

FIGURE 2. Relative frequency for each item searched (A), mean number of items per article (i.e., application of a multivariable regression model, stratified by impact factor and by sample size) (B), study design, (C) and type of model used (D).

We are aware that the items we reviewed need not have the same relevance and weight—and are not even always necessary. Assessing their interactions is not always justified, especially in small samples, and there are studies on medical interventions in which confounding could be considered negligible. Furthermore, there are other important aspects that would affect the quality of the analysis and results (e.g., the model was prespecified prior to undertaking the data analysis, or the research team had insufficient statistical background and knowledge).

Finally, our review was not paired and there could be a certain interrater variability. We attempted to minimize this potential limitation in 2 ways: very detailed specification of each item to be reviewed, all of which were easily identifiable; and prior training of reviewers with a pilot test. In addition, testing for agreement after completing the review showed a high level of intra- and interrater agreement (Supplementary Table S2, http://links.lww.com/MD/A979).

CONCLUSIONS

Statistical adjustment using MRM is a powerful tool for isolating the actual effect of exposure factors on potential confounders. However, the use of these techniques is not free of potential errors because they have strong underlying assumptions that must be tested. Our study showed that, despite the availability of known statistical tools that allow the evaluation of how well the models meet the conditions for their application, only a troublingly low percentage of published articles report information about model validation or measures to ensure the rigorous application of MRMs as an adjustment method. Given the importance of these statistical methods to the final conclusions, biomedical journals should require greater rigor in reporting the assumptions of the MRMs in the methods and results of observational studies.

ACKNOWLEDGMENTS

The authors thank Lluís Alvarez for his role in the management of full manuscripts; Gisela Galindo, and Inés Cruz for their valuable advice on the latest versions; and Elaine Lilly for English editing, supported by IDIAP-Jordi Gol.

REFERENCES

1. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet.* 2002;359:248–252.
2. Bender R. Introduction to the use of regression models in epidemiology. In: Mukesh Verma, ed. *Methods in Molecular Biology, Cancer Epidemiology. Vol. 471.* United States: Springer Science; 2009:179–195.

3. Yergens DW, Dutton DJ, Patten SB. An overview of the statistical methods reported by studies using the Canadian community health survey. *BMC Med Res Methodol.* 2014;14:1.
4. Groenwold RH, Van Deursen AM, Hoes AW, et al. Poor quality of reporting confounding bias in observational intervention studies: a systematic review. *Ann Epidemiol.* 2008;18:746–751.
5. Gentle JE, Härdle WK, Mori Y. How computational statistics became the backbone of modern data science. In: Handbook of Computational Statistics. Berlin Heidelberg: Springer; 2012:3–16.
6. Dobson AJ. An Introduction to Generalized Linear Models 2nd ed. United States of America: Chapman and Hall; 2001.
7. Gerhard T. Bias: considerations for research practice. *Am J Health Syst Pharm.* 2008;65:2159–2168.
8. Liang W, Zhao Y, Lee AH. An investigation of the significance of residual confounding effect. *Biomed Res Int.* 2014;2014:658056.
9. Groenwold RH, Klungel OH, Altman DG, et al. Adjustment for continuous confounders: an example of how to prevent residual confounding. *CMAJ.* 2013;185:401–406.
10. Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Stat Med.* 1995;14:1707–1723.
11. Vittinghoff E, Shiboski S, McCulloch CE. Regression Methods in Biostatistics.Springer; 2005.
12. Oxman AD. GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ.* 2004;328:1490–1494.
13. Wu R, Glen P, Ramsay T, et al. Reporting quality of statistical methods in surgical observational studies: protocol for systematic review. *Syst Rev.* 2014;3:70.
14. Casals M, Girabent-Farres M, Carrasco JL. Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000–2012): a systematic review. *PLoS One.* 2014;9:e112653.
15. Strasak AM, Zaman Q, Pfeiffer KP, et al. Statistical errors in medical research-a review of common pitfalls. *Swiss Med Wkly.* 2007;137 (3/4):44.
16. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ.* 2009;338:b605.
17. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* John Wiley & Sons; 2004. http://books.google.es/books?id=Po0RLQ7USIMC. Accessed June 4, 2015.
18. Müllner M, Matthews H, Altman D. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. *Ann Intern Med.* 2002;136:122–126.
19. Real J, Cleries R, Forne C, et al. Use of multiple regression models in observational studies (1970–2013) and requirements of the STROBE guidelines in Spanish scientific journals. *Semergen.* 2015;11:S1138–S3593.
20. Lee YH, Hsu CY, Hsia CY, et al. A prognostic model for patients with hepatocellular carcinoma within the Milan criteria undergoing non-transplant therapies, based on 1106 patients. *Aliment Pharmacol Ther.* 2012;36:551–559.
21. Chen S, Huang L, Liu Y, et al. The predictive and prognostic significance of pre- and post-treatment topoisomerase IIalpha in anthracycline-based neoadjuvant chemotherapy for local advanced breast cancer. *Eur J Surg Oncol.* 2013;39:619–626.
22. Moons KG, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ.* 2009;338:b606.
23. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35:1925–1931.
24. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ.* 2015;350:g7594.
25. Schulz KF, Altman DG, Moher D. CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med.* 2010;152:726–732.
26. Noah N. The STROBE initiative: STrengthening the reporting of OBservational studies in epidemiology (STROBE). *Epidemiol Infect.* 2008;136:865.
27. Lang T, Altman D. Statistical analyses and methods in the published literature: The SAMPL guidelines. *Guidelines for Reporting Health Research.* 2014:264–274.
28. Evans R. AMA manual of style-A guide for authors and editors. *Nurs Stand.* 2007;21:31–131.
29. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol.* 2007;166:646–655.
30. Ho KM. Residual confounding in observational studies. *Anesthesiology.* 2009;110:430author reply 430.
31. Sainani K. The limitations of statistical adjustment. *PM R.* 2011;3:868–872.
32. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Stat Med.* 2013;32:67–80.
33. Ho DE, Imai K, King G, et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal.* 2007;15:199–236.
34. Benedetti A, Abrahamowicz M. Using generalized additive models to reduce residual confounding. *Stat Med.* 2004;23:3781–3801.