Errors in the use of Multivariable Logistic Regression Analysis: An Empirical Analysis

Sir,

The issue (October 2019–December 2019) of the Indian Journal of Community Medicine published 17 original articles, and of these, five studies applied the multivariable logistic regression (MLR), and one study applied the multinomial logistic regression.[1-6] The MLR is widely applied statistical methods in the medical journals to assess the magnitude of association between binary outcomes and sets of independent qualitative and quantitative variables. I would like to highlight common errors that occurred in MLR articles published in this issue. In addition, some other statistical issues found in these articles are also be highlighted. The aim of this letter is to improve the quality of MLR in medical journals and not to hurt any author of these articles. These errors divided under the following major heading:

Selection of potential variables and mismatch between selecting criterion and actual included variable in the MLR analysis: univariable analysis (UA) is frequently applied mode for selecting the variables for MLR, and cutoff *P* value for candidate variables varies from study to study, the most commonly found cutoff for $P < 0.05$. This cutoff usually not recommended to select the variables because it may be possible that a variable individually insignificant but in multivariable setup is significant or vice versa. Thus, statistician recommended to inflate the cutoff *P* value of UA to 0.2 or 0.25 to minimize this. Two studies chosen the cutoff of < 0.2 or < 0.25,[1,3] whereas four studies selected < 0.05.[2,4-6] Four studies had a mismatch in variable selecting criterion and actually included in the MLR as follows:

(a) In Dubey *et al*., four variables such as gender, age, vacation, and the number of siblings selected with $P < 0.2$ on UA but included only first three in MLR,[1] (b) Madasu *et al*. selected the variables with cutoff $P < 0.25$ in UA but included the age having $P = 0.74$ and excluded the family type having $P = 0.17$,[2] (c) Rampur *et al*. chose $P < 0.05$ to select the variables on the basis of *t*-test, the variable tangible support score showed $P > 0.05$ but included in the MLR model, and[3] (d) Ram *et al*. reported to include the statistically significant variables (<0.05) on UA but included dowry variable with $P = 0.068$ in MLR.[5]

Interpretation of odds ratio as relative risk ratio

MLR results are not directly interpretable, either probabilities or relative risk ratio (RR) unless the outcome of interest is rare (<10%). Interpretation of odds ratio (OR) for common outcomes as RR overestimates the association effect, and overestimation increases as the prevalence of outcome increases when OR > 1. The prevalence of alcohol relapsed is around 55.4%, which was not rare, and the author reported the OR as RR in their article, "People with a high level of craving have 1.8 times chance of relapse as compared to people with a low craving."[3] The reported OR was 1.78 (95% confidence interval [CI]: 1.25–2.54). The correct interpretation is "Odds of alcoholic relapse have a high level of craving is 1.8 times more than odds of alcoholic relapse having a low level of craving." In Madasu *et al*. reported that "female sex was found to be two times more associated with anxiety disorder than male."[2]

Wrong interpretation of word adjusted

The word adjusted in MLR means the risk factor of interest to be adjusted for other variables included in the MLR. Another explanation, each variable to be adjusted for all other variables presents in the final MLR model. The variables removed on the basis of univariable screening should not be included in the list of adjustments. Enter the method applied to assess the independent effect, and this method retained all the variables included in MLR initially whether it is significant or not. Automatic methods such as forward, backward, or stepwise that add/remove variables according to the specified criteria and produce the final model containing only variables that fulfilled the criteria. The word adjustment is used only in the context of variables left in the final model and not for variables initially considered in the MLR model. Dubey *et al*. reported that "OR for poor sleep quality across various characteristics was achieved using logistic regression after adjusting age, gender number of sibling, and vacation."[1] This interpretation represents that author had tested other characteristic effects after adjusting these four characteristics using MLR. In fact, they considered only three variables: age, gender, and vacation in MLR. Rampure *et al*. wrote in the statistical section, "MLR was done to determine the independent factors associated with alcohol relapse and to adjust for confounders."[3] It is unclear from the table and text about factors and confounders variables.

Nonreporting of 95% CI and contradiction between reported 95% CI of OR and *P* value: The main objective of inferential statistics is to estimate the errors, which allows us to understand what would happen when the study repeats multiple number of times, and 95% of CI helps in determining this. The researcher sometimes deliberately hides the 95% CI due to the wide interval. Rao *et al*. reported the OR without giving the 95% CI.[4]

Inclusion of one in 95% CI of OR indicates a nonstatistically significant association between the outcome of interest and respective variable at 5% level of significance. In Madasu *et al*., the 95% CI for chronic disease OR includes 0.93–9.85

and respective *P* value = 0.03. Recalculation of OR values found OR = 3.07 and 95% CI (1.10–8.61) with *P* = 0.03, which did not include one.[2] Rampur *et al*. reported OR of social support as 0.33 (95% CI: 0.14–0.79); with *P* = 0.13. This 95% CI did not include one showed significant association with the outcome while *P* value > 0.05.[3]

Change in scale of a variable from univariable to multivariable

The dichotomization of continuous variables loses one-third of the information. Various methods have been now proposed in the literature to include continuous variables in MLR. The change of scale from UA to MLR affects the significance of variable as well as its interpretation. Albeit, it may be possible a univariably significant continuous variable may not be significant even in univariably when dichotomized. Rampur *et al*. dichotomous all the continuous score as high and low without proper justification and select the variables for MLR on the basis of Student's *t*-test significance.[3]

Inclusion of co-linear variables

The validity of MLR depends on the number and suitability of variables included in the model. When included variables convey closely related information leads to great uncertainty in the estimation effect of these variables on the outcome of interest. Ram *et al*. include socioeconomic status and occupation of women who convey the overlapping effect on the outcome of interest.[5]

Wrong coding

Statistical software by default considers code 0 as a reference category. Wrong selection of reference code makes risk factors as protective factors and vice versa. Dubey *et al*. OR of poor sleep for vacation (yes) is reported as 1.5 (0.8–2.9) showed risk factors while it is protective with OR as 0.68 (0.34–1.33).[1] Similarly, in Madasu *et al*., the OR for anxiety disorder for schooling status (out of school) should be 1.29 (0.76–2.21) instead of 0.77 (0.45–1.32).[2]

Sample size

The sample size calculation detail should be clear and complete so that it can be reproducible by readers. The sample size is a common issue in these articles. All the articles were cross-sectional studies with the primary objective to assess the prevalence and apply standard prevalence formula to determine the sample size. None of the articles adjusted the sample size for MLR. Inadequate sample size leads to estimate untrustworthy parameter value, and inflate the standard error causes wide 95% CIs.

Adequate analysis and complete reporting are required for the primary outcome variable on which sample size is determined. For example, estimation as primary outcome should be reported with its 95% CI, whereas hypothesis testing outcome should be reported along with suitable statistical test and its *P* value or effect size with 95% CI. Madasu *et al*. calculated sample size to compare the proportion of anxiety disorder among adolescents with literature proportion of anxiety disorder (hypothesis testing), whereas the primary objective was to estimate the prevalence of anxiety disorder.[2] Rao *et al.* reported "a pilot study was conducted, and the sample size was calculated as 48" without providing any further information. It looks like a confusing statement.[4]

Other statistical issues

(a) The number and percentage calculations were wrong for most of the variables.[2] (b) Maity *et al*. did not report the magnitude of the regression coefficient and OR of multinomial logistic regression.[6] (c) Rampur *et al*. reported the *t*-statistic of desirable events in relapse and abstinence subject as 1.96 and corresponding *P* value as 0.01, and similarly, *t*-statistic as − 2.10 with *P* = 0.90 for undesirable events. Both these *P* value are not matching with *t*-statistic.[3]

Quality of MLR published in high impact factor Indian medical journals on the basis of 10-point well-established criteria revealed that MLR quality needs improvement in all its dimensions.[7] Regression models require in-depth understanding of model building, it associated assumptions, adequate and complete reporting of results, and correct interpretation to get correct and reliable conclusions from the model results. Authors become more acquainted with MLR from designing to reporting or consult competent statisticians while applying the MLR. Interested authors can find cautions needed in planning, analysis, and reporting of MLR.[8]

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

Rajeev Kumar Malhotra

Delhi Cancer Registry, Dr. BRA IRCH, AIIMS, Delhi, India

Address for correspondence: Dr. Rajeev Kumar Malhotra, Room No. 24, Delhi Cancer Registry, Dr. BRA IRCH, AIIMS, Delhi - 110 029, India.
E-mail: rejeev.Kumar.malhotra@gmail.com

References

1. Dubey M, Nongkynrih B, Gupta SK, Kalaivani M, Goswami AK, Salve HR. Sleep quality assessment of adolescents residing in an urban resettlement colony, New Delhi, India. Indian J Community Med 2019;44:271-6.
2. Madasu S, Malhotra S, Kant S, Sagar R, Mishra AK, Misra P, *et al*. Anxiety disorders among adolescents in a rural area of northern India using screen for child anxiety-related emotional disorders tools: A community-based study. Indian J Community Med 2019;44:317-21.
3. Rampure R, Inbaraj LR, Elizabeth CG, Norman G. Factors contributing to alcohol relapse in a rural population: lessons from a camp-based de-addiction model from rural Karnataka. Indian J Community Med 2019;44:307-12.
4. Rao KA, Thomas S, Kumar JK, Narayan V. Prevalence of dentinal hypersensitivity and dental erosion among competitive swimmers, Kerala, India. Indian J Community Med 2019;44:390-3.
5. Ram A, Victor CP, Christy H, Hembrom S, Cherian AG, Mohan VR. Domestic violence and its determinants among 15-49-year-old women in a rural block in South India. India J Community Med 2019;44:362-7.
6. Maity B, Chaudhuri D, Saha I, Sen M. Association of nutritional

status with depression and cognitive function of older women residing in old-age homes of Kolkata, India. Indian J Community Med 2019;44:328-31.
7. Kumar A, Indrayan A, Chhabra P. Reporting quality of multivariable logistic regression in selected Indian journals. J Postgrad Med 2012;58:123-6.
8. Kumar R, Chhabra P. Caution requires during planning, analysis and reporting of multivariable logistic regression. Curr Med Res Pract 2014;4:31-9.

Access this article online

Quick Response Code:

Website: www.ijcm.org.in

DOI: 10.4103/ijcm.IJCM_16_20

How to cite this article: Malhotra RK. Errors in the use of multivariable logistic regression analysis: An empirical analysis. Indian J Community Med 2020;45:560-2.

Received: 08-01-20, **Accepted:** 04-06-20, **Published:** 28-10-20