Ophthalmic Statistics Note 12: Multivariable or Multivariate: what's in a name?

Catey Bunce 1, Gabriela Czanner 2, Mariusz Grzeda 3, Caroline J Doré 4, Nick Freemantle 5

*Corresponding author

1 Department of Primary Care & Public Health, Kings College London, London, SE1 1UL, UK

Email : catey.bunce@kcl.ac.uk

2 University of Liverpool, Liverpool, Merseyside, L69 3BX, UK

3 School of Social & Community Medicine, University of Bristol, Bristol, UK

4 Comprehensive Clinical Trials Unit, University College London, Great Britain

5 Medical School, University College London, Great Britain

Keywords: medical statistics, multivariate, multivariable

Word count: [2178] (including tables and references)

A senior colleague asks me to critique a paper which reports to have used multivariate statistical methods to suggest an inhibitory effect of maternal smoking on the development of severe retinopathy of prematurity (ROP). [1] S/he is concerned by the paper because the abstract suggests a positive effect of maternal smoking which flies very much against public health messages in general regarding smoking but is reassured by the fact that complex statistical methods – namely multivariate techniques have been employed.

I access the internet and find that the paper has been published in a peer-reviewed journal of high repute and that it reports an analysis conducted using data from 86 premature (< 32 weeks' gestation) infants. ROP grading had been evaluated in accordance with the International Classification of Retinopathy of Prematurity. [2] The authors explored clinical characteristics associated with the proportions of babies who had developed severe ROP (defined as stage 3 with plus disease). Several characteristics had been recorded for each baby or mother – including birth weight, gestational age, gender of the baby, oxygen supplementation and maternal smoking. The authors report results of both univariate and multivariate logistic regression analyses and that analyses were conducted using STATA version 10 and R version 2.71.[3, 4]

I am not familiar with the term multivariate and so I consult the internet and statistical books.[5, 6, 7] I learn that multivariate techniques are very different to univariate techniques. I learn that the term "multivariate" in general means "many variables" but in statistical jargon it has come to have a more specific meaning; many dependent (response) variables or alternatively variables where there is no

hierarchy – ie variables are not classified into response and predictors but are regarded as being on an equal footing. [8]

In univariate techniques, there is a single outcome or dependent (response) variable (in this instance development of severe ROP) and one independent or explanatory variable which may sometimes also be termed 'covariate' (in this instance birthweight or gestational age etc). Univariate logistic regression could be used to identify which variables are associated with the odds of severe ROP. I would create a series of models for each of the explanatory variables that have been recorded within this study and in each case I would explore the association between that explanatory variable and severe ROP development. Whilst this might be of interest, we would probably be interested to use information on several of the recorded variables simultaneously to determine disease development, and for this we would use multiple variable or multivariable logistic regression. Multivariable methods, are the tools to use when there is one dependent/response variable but more than one independent/predictor/explanatory variable. Multivariable methods may be used to identify which of several potential predictor variables is "important" , to develop a prognostic model from several predictor variables or to remove the possible effects of "nuisance" variables or confounders.

Whilst we understand univariate and multivariable techniques, I am unsure whether multivariate is the same as multivariable. As mentioned above, the statistical jargon implies that multivariate is only to be used in situations where all variables are treated equally or there is more than one dependent variable.

The first situation ie where variables are regarded as on an equal footing covers methods that are typically used for data reduction purposes (ie reducing the number of variables in an analysis), to examine the relationships between individuals or the relationships between variables or to develop rules to classify subjects into groups. Such methods might also be used in the development of measures or scales for complex underlying concepts such as 'visual dysfunction' or 'vision disability'. [9, 10] Concepts of this type are quite abstract and for this reason they are frequently called 'latent variables' that can only be represented meaningfully by combining several observable variables (sometimes also called manifest indicators of latent variables).[11, 12] This type of multivariate analysis includes methods such as principal component analysis , factor analysis , item response theory, latent class analysis, linear discriminant analysis, multidimensional scaling and many others.[7, 8]

The second situation is where we have two or more dependent variables and we wish to examine relationships between these and several explanatory variables. This class of multivariate analyses includes multivariate linear regression and multivariate analysis of variance.

It appears to me that multivariate methods have not been employed. This paper has a single dependent variable and multiple independent variables and the authors have used multivariable logistic regression and not a multivariate method. I wonder whether it matters and learn that the statistical methods section within a paper should allow the reader to comprehend fully what has been done.[13] The abstract of this paper suggests something very different to what has been done.

I advise my senior colleague that multivariate methods have not been used and that perhaps this misunderstanding throws doubt upon the statistical validity. [14] My colleague looks at the paper and says that whilst there may have been an error in the description of the methods, this doesn't necessarily mean that the conclusions are incorrect. He comments that the authors have clearly described the statistical packages that they have used and that a robust classification system has been used to determine retinopathy. S/he asks me to prepare a critique of the paper for a journal

club meeting and to try to determine whether there is robust support for the assertion that maternal smoking might have an inhibitory effect on development of severe ROP.

I consider other aspects of the multivariable model – my understanding is that the regression coefficients provided no longer give me a simple assessment of how that factor relates to the outcome variable but something more complicated. In a model with two independent variables or covariates (say maternal smoking and gestational age), the coefficients, now called marginal (or adjusted or conditional) coefficients, provide an estimate of the effect of maternal smoking on development of ROP whilst "holding" gestational age constant. My understanding of this is that it therefore gives me a measure of association between the odds of severe ROP and maternal smoking in babies with similar gestational ages - for example, two babies with a gestational age of 27 weeks or two babies with a gestational age of 30 weeks. If only these two covariates are in the model, an assumption is being made that the effect of maternal smoking on the odds of severe ROP is the same irrespective of gestational age. If the effect of maternal smoking differed according to the gestational age of the baby (older babies having been exposed to indirect smoking for longer than younger babies), I learn that an interaction term would need to be included in the model. There is no mention within the paper of an examination of the potential for interaction but I learn that interactions are often not explored fully because detecting them requires a lot of data and frequently there is insufficient data to fully explore these.

In a model with three independent variables or covariates (say maternal smoking, gestational age and the gender of the baby) the marginal coefficients are giving an estimate of the association between the odds of severe ROP and maternal smoking whilst "holding" gestational age and the gender of the baby constant. The model is therefore looking at the effect of maternal smoking versus not smoking in babies of the same sex and of the same gestational age. Again, this model is making an assumption that these covariate effects are not dependent on the levels of the other factors – ie that there are no interactions. Whilst 86 premature babies seemed like a reasonable number to explore associations, I now see why large numbers are needed to assess models reliably. The more variables that are included in the model the greater the data are stretched and there simply will be no data to support the examination. A model is being fitted with limited ability to assess its fit.

The multivariable model reported in the paper contains 7 covariates. I learn that a rule of thumb for logistic regression models is that the number of observed events (or non-events, whichever is smaller) for each independent variable considered within a model should be at least ten. [16] Ideally therefore, given that there were only 27 babies who developed severe ROP, fewer than three independent variables should have been considered for inclusion in the model, rather than at least ten independent variables examined by the authors.

In the model with three factors, the logistic regression model gives me an estimate of the effect of maternal smoking on severe ROP in premature babies of the same gestational age and the same gender. Each time a variable is added to the model, I must consider that an additional variable is being held constant. I start to realise how little data are contributing to these adjusted estimates. For example, only one mother of the 27 babies with severe retinopathy was a smoker.

The odds ratio estimate under scrutiny is 0.01 with a confidence interval of 0.00 to 0.48. If this were a univariate model, the interpretation would be that the odds of severe ROP in a premature baby of a mother who smoked is 0.01 times that of the odds of severe ROP in a premature baby of a mother who had not smoked. This, however, is a multivariable model and so I now acknowledge that it is

actually saying something slightly different – ie that *if* all of the other covariates in the model are held constant, this would be the effect of maternal smoking.

In addition to statistical uncertainty I determine that there are sources of bias that do not appear to have been adequately dealt with. Smoking status was self-reported by mothers at their first visit to the mother and baby centre - might some wish not to disclose such information for fear of recrimination? Mothers who discontinued smoking during pregnancy or who had an "uncertain" smoking status were excluded from the study (selection bias) and we are not told how many such exclusions there were.

At the journal club I present the paper. We conclude that there is no robust information provided suggesting an inhibitory effect of maternal smoking on the development of severe ROP and have a greater understanding of multivariate and multivariable statistical techniques.

Lessons learned

Multivariate methods are not the same as multivariable methods.

Multivariate methods have more than one dependent variable or place variables on an equal footing.

Multivariable methods have one dependent variable and more than one independent variables or covariates.

Regression coefficients from multivariable models need careful interpretation as their meaning differs to that from a univariate model.

The number of observed events (or non-events, whichever is smaller) for each independent variable considered within a multiple variable logistic regression model should be at least ten.

Contributors

CB drafted the paper. GZ, MG, CJD, CB and NF critically reviewed and revised the paper.

Funding

CB is part funded/supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests

None declared.

1 Hirabayashi H, Honda S, Morioka I, Yokoyama N, Sugiyama D, Nishimura K, Matsuo M, Negi A.

Inhibitory effects of maternal smoking on the development of severe retinopathy of prematurity. Eye (Lond). 2010 Jun;24(6):1024-7. doi: 10.1038/eye.2009.263. Epub 2009 Nov 6.

2. The International Classification of Retinopathy of Prematurity revisited. International Committee for the Classification of Retinopathy of Prematurity. Arch Ophthalmol. 2005 Jul;123(7):991-9. Review.

3. STATA version 10, StataCorp, Lakeway Drive College Station, Texas, USA

4. R version 2.71 (R foundation for Statistical Computing, Vienna, Austria)

5. Tabachnick BG, Fidell LS. Using Multivariate Statistics. Boston: Pearson/Allyn & Bacon, 2007

6. Everitt B. Statistical Methods in Medical Investigations. London: E. Arnold, 1994

7. Chatfield C, Collins A J. Introduction to Multivariate Analysis. Chapman & Hall, London and New York 1985.

8 Streiner DL. An introduction to multivariate statistics. Can J Psychiatry. 1993;38(1):9-13

9 Donovan J et al. The development and validation of a questionnaire to assess visual symptoms/dysfunction and impact on quality of life in cataract patients: the Visual Symptoms and Quality of life (VSQ) Questionnaire, Ophthalmic Epidemiology. 2003;10(1):49–65

10 Massof R. The measurement of vision disability, Optometry Vision Science. 2002; 79(8):516-52.

11 Wilson M. (2005). Constructing Measures. Mahawah, New Jersey; London: Lawrence Erlbaum Associates, Publishers

12 Bollen KA. Latent variables in psychology and the social sciences, Annual Review of Psychology 2002. 53:605–34

13. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. Int J Nurs Stud. 2015;52(1):5-9

14. Hidalgo B, Goodman M. Multivariate or multivariable regression? Am J Public Health. 2013 Jan;103(1):39-40. doi: 10.2105/AJPH.2012.300897. Epub 2012 Nov 15. Review

15. Peduzzi P1, 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 Dec;49(12):1373-9.