LETTERS

ACHIEVING CONSENSUS ON TERMINOLOGY DESCRIBING MULTIVARIABLE ANALYSES

In their recent article, Hidalgo and Goodman[1] call our attention to the need for consistent and distinctive use of the terms "multivariable" and "multivariate." They introduced a point of confusion, however, with their suggestion that the terms "linear, logistic, multivariate, or proportional hazards" be employed to indicate continuous, dichotomous, repeated measures, or time-to-event outcomes, respectively. I find their suggestion confusing because it suggests the absence of an overlap between "linear," "logistic," and "multivariate." Yet a regression model fit to repeated-measures data may assume a normal *or* logistic distribution (or any of a number of other distributions), making it a multivariate linear or multivariate logistic regression model.

I believe their article invites two additional teaching points for reinforcement, which I underline here. I surveyed 22 empirical articles published in the same January 2013 issue. Of these, three articles (13.6%) used the term "multivariate" incorrectly, including one article that used the term "bivariate." Five articles (22.7%) used "multivariate," "multiple" (i.e., multiple regression), and "multivariable" interchangeably, including one article that used the term "bivariate." Three articles (13.6%) used the term "multivariate" correctly in the context of repeated-measures or nested data, while eleven (50%) contained no violations.

First, the term "univariate" is most appropriate (and perhaps is unnecessarily described explicitly as such) when there is only one response variable per observation. Depending on whether there is one explanatory variable or multiple explanatory variables, the terms "univariable" and "multivariable" (i.e., multiple) would help to additionally clarify the kind of univariate analysis being conducted. A *t*-test comparing mean levels of a response variable between two subgroups is a univariable analysis, and so is a regression model of the same response variable with the subgroup specified as the single binary explanatory variable. Use of the term "bivariate" to describe such a *t*-test, while common (and observed twice in the cursory survey described above), introduces unnecessary confusion and should be discouraged.

Second, the term "multivariate" should be understood to apply to a diverse set of methods that allow for more than one response per observation.[2] Hidalgo and Goodman noted certain applications of repeated measures regression, or—to retain consistency with the terminology I elaborated upon—multivariate multivariable regression. This presents a compelling rationale for why the terms "multivariate" and "multivariable" should not be used interchangeably. Other types of statistical analyses are also classified as "multivariate," including discriminant analysis, canonical correlation, and principal components analysis.

The nuances in the use of statistical terminology described by Hidalgo and Goodman have not gained formal traction at most peer-reviewed journals.[3] This is likely because equally reasonable perspectives are also taught. For example, in one leading textbook for clinical practitioners, the author says that "multivariate analysis refers to simultaneously predicting multiple outcomes."[4(p1)] But the author also writes (contrary to the recommendation above),

> I think it is more informative to restrict the term "univariate" to analyses of a single variable, while restricting the term "bivariate" to refer to the association between two variables.[4(p5)]

Ultimately, achieving consensus on these issues will help to avoid further confusion and facilitate substantive progress on communicating the results of public health research in the published literature. ■

Alexander C. Tsai, MD, PhD

About the Author

Alexander C. Tsai is with the Center for Global Health and the Chester M. Pierce, MD Division of Global Psychiatry, Massachusetts General Hospital, Boston. He is also with Harvard Medical School, Boston, MA.

Correspondence should be sent to Alexander Tsai, MD, 100 Cambridge Street, 15th floor, Boston, MA 02114 (e-mail: actsai@partners.org). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This letter was accepted January 3, 2013.
doi:10.2105/AJPH.2013.301234

Acknowledgments

A. C. Tsai acknowledges salary support from the National Institutes of Health (NIH K23 Mentored Patient-Oriented Research Career Development Award MH-096620).

References

1. Hidalgo B, Goodman M. Multivariate or multivariable regression? *Am J Public Health.* 2013;103(1):39–40.
2. Rencher AC. *Methods of Multivariate Analysis.* 2nd ed. New York, NY: Wiley-Interscience; 2002.
3. Peters TJ. Multifarious terminology: multivariable or multivariate? univariable or univariate? *Paediatr Perinat Epidemiol.* 2008;22(6):506.
4. Katz MH. *Multivariable Analysis: A Practice Guide for Clinicians.* New York, NY: Cambridge University Press; 1999.

Letters to the editor referring to a recent Journal article are encouraged up to 3 months after the article's appearance. By submitting a letter to the editor, the author gives permission for its publication in the Journal. Letters should not duplicate material being published or submitted elsewhere. The editors reserve the right to edit and abridge letters and to publish responses.

Text is limited to 400 words and 10 references. Submit online at www.editorialmanager.com/ajph for immediate Web posting, or at ajph.edmgr.com for later print publication. Online responses are automatically considered for print publication. Queries should be addressed to the Editor-in-Chief, Mary E. Northridge, PhD, MPH, at men6@nyu.edu.

HIDALGO AND GOODMAN RESPOND

We appreciate Tsai's letter, and we acknowledge that, in our recent article, we provided an oversimplification of a complex topic. We did so to present a clear argument to those

without formal technical training in statistics. We apologize for any confusion our format may have caused, and we would like to take this opportunity to address Tsai's points as well as to provide further clarification.

We agree that multivariate statistics is a broad area,[1] but the purpose of our article was to encourage authors to use the term "multivariate" correctly and not to use it when the models they describe are really multivariable. The most important point to consider is that we are discussing the nuance in the use of terminology in the context of regression only, and not in the broader context of the type of statistical analysis conducted. We believe that the terms "univariable" and "multivariable" should only be used in the regression context to describe the number of predictors in the model, whereas the terms "univariate" (1 variable), "bivariate" (2 variables), and "multivariate" (multiple variables) should be used to describe the type of statistical analysis being conducted.[2] Thus a *t*-test is a type of bivariate statistical analysis because of the use of two variables.[3]

We concur that multivariate models can be linear, logistic, or proportional hazards, and we did not mean to suggest that these were mutually exclusive. These terms are simply how we think regression models should be described in the public health literature. For example, if authors use the terms "linear" or "logistic regression," then "univariate" is implied. But they should also specify whether the model is simple or multivariable. If they use a multivariate regression model, they should still specify the type of model (e.g., linear, logistic, or proportional hazards) and whether it is unadjusted (simple) or adjusted (multivariable). We appreciate Tsai's affirmation that the terms "multivariate" and "multivariable" should not be used interchangeably as they have two distinct meanings and that some regression models would be most appropriately defined as multivariate multivariable models.

Until we reach a consensus on how to describe regression models, we encourage readers to make sure they understand what is meant when the term "multivariate" is used to define a regression model. ■

Bertha Hidalgo, PhD, MPH
Melody Goodman, PhD, MS

About the Authors

Bertha Hidalgo is with the Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham. Melody Goodman is with the Department of Surgery, Division of Public Health Sciences, School of Medicine, Washington University in St. Louis, St. Louis, MO.

Correspondence should be sent to Bertha Hidalgo, PhD, MPH, 1665 University Blvd., RPHB 443, Birmingham, AL 35226 (e-mail: bhidalgo@uab.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This letter was accepted January 16, 2013.
doi:10.2105/AJPH.2013.301245

Contributors

Both authors contributed equally to this letter.

Acknowledgments

B. Hidalgo was supported in part by a postdoctoral training grant from the National Heart, Lung, and Blood Institute (grant T32HL072757). M. Goodman was supported by the Siteman Cancer Center, the National Cancer Institute (grant U54CA153460), and the Washington University School of Medicine Faculty Diversity Scholars Program.

References

1. Marcoulides GA, Hershberger SL. *Multivariate Statistical Methods: A First Course.* Mahwah, NJ: Lawrence Erlbaum Associates; 1997.

2. Diamantopoulos A, Schlegelmich BB. *Taking the Fear Out of Data Analysis: A Step-by-Step Approach.* London, UK: Thomson Learning; 1997.

3. Sims RL. *Bivariate Data Analysis: A Practical Guide.* Hauppauge, NY: Nova Science Publishers; 2004.