routine use. Now it is time to let the volumes come, filled with important questions, answers and interactive online discussion among researchers, clinicians, patients, educators and policy makers—and to turn the volume up so decision makers in both clinical and policy settings hear the messages.

To read or post commentaries in response to this article, see it online at http://www.annfammed.org/cgi/content/full/2/3/197.

Key words: Medical writing; medical publishing; primary care; family medicine; health policy

Submitted April 26, 2004; accepted April 29, 2004.

References

1. Donaldson MS, Yordy KD, Lohr KN, Vanselow NA, eds. Committee on the Future of Primary Care Services. Institute of Medicine. *Primary Care: America's Health in a New Era.* Washington, DC: National Academy Press; 1996.
2. Institute of Medicine. Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Institute of Medicine. Washington DC: National Academies Press; 2001.
3. Corrigan JM, Greiner A, Erickson SM, eds. Committee on Rapid Advance Demonstration Projects: *Health Care Finance and Delivery Systems.* Washington, DC: National Academies Press; 2003
5. Future of Family Medicine Project Leadership Committee. The future of family medicine: a collaborative project of the family medicine community. *Ann Fam Med.* 2004;2(Suppl 1):S3-S32.

EDITORIAL

On the Nature and Analysis of Clustered Data

Stephen J. Zyzanski, PhD[1]
Susan A. Flocke, PhD[1]
L. Miriam Dickinson, PhD[2]

[1]Departments of Family Medicine, Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio

[2]Department of Family Medicine, University of Colorado Health Science Center, Denver, Colo

Ann Fam Med 2004;2:199-200. DOI: 10.1370/afm.197.

Studies in which data from multiple patients are collected per clinician or per practice are becoming common in primary care research, particularly with the increase of studies conducted in practice-based research networks. These studies generate data that are clustered. A special case of clustered data is an intervention study where clinicians or practices are randomized into an intervention or control group. In such cluster-randomized designs, all patients of a clinician or practice are assigned to the same treatment, and this design is often used when logistics of implementation or the need to avoid contamination of treatment arms is a priority.

A major issue in the analysis of clustered data is that observations within a cluster are not independent, and the degree of similarity is typically measured by the intracluster correlation coefficient (ICC).[1] Ignoring the intracluster correlation in the analysis could lead to incorrect *P* values, confidence intervals that are too small, and biased estimates and effect sizes, all of which can lead to incorrect interpretation of associations between variables.[2] Failure to take into account the clustered structure of the study design during the planning phase of the study also can lead to underpowered study designs in which the effective sample size and statistical power to detect differences are smaller than planned.

In most situations, the numeric value of the intracluster correlation tends to be small and positive. Several authors have provided guidelines for interpreting the magnitude of the intraclass correlation[3] with small, medium, and large values of the intraclass correlation coefficients reported as .05, .10, and .15. Small values of the intracluster correlation can be deceiving, however. Investigators need to be aware that the cluster effect is a combination of both the intracluster correlation and the cluster size. Small intracluster correlations

Conflicts of interest: none reported

CORRESPONDING AUTHOR

Stephen J. Zyzanski, PhD
Department of Family Medicine
Case Western Reserve University
11001 Cedar Ave, Suite 306
Cleveland, OH 44106
sjz@po.cwru.edu

coupled with large cluster size can still affect the validity of conventional statistical analyses.

Although clustered data are common, investigators often overlook both the special analysis challenges and the unique opportunities inherent with clustered data.[4,5] In this issue of the *Annals*, Reed suggests a convenient correction procedure to address clustered data.[6] The correction involves applying a formula to the standard errors and then conducting the planned analysis with the corrected standard errors. Also in this issue, the article by Killip et al[7] provides a formula to compute an effective sample size for clustered data. Computation of the effective sample size is important, as it avoids costly sample size errors caused by underpowered studies. Examples in the Killip et al article show how the intracluster correlation, number of observations within a cluster, and number of clusters are all interrelated in estimating sample size and power for clustered data.

Clustered data imply a hierarchical nature to the data, and while many levels can be considered, two levels are most commonly specified. The outcome measure is always assessed at the lowest level. Explanatory variables, however, may be considered at any of the levels (eg, patient variables and/or physician or practice level variables). Consequently, clustered data provide considerable opportunities to explore, in greater depth, the interrelationships among variables at any level; these analyses are generically called multilevel analyses.

Considering an example of data with patients clustered with physicians, a comprehensive multilevel data analysis aims to assess the direct effect of patient and clinician/practice level variables on the outcome. One could also determine whether the variables at the clinician/practice level serve as moderators of patient level relationships by testing cross-level interactions between variables from the patient level and the physician level.[8] Hence, multilevel analyses are designed to analyze variables from different levels simultaneously, all the while taking into account the intracluster correlation.

Statistical software to conduct these types of analyses and for computing sample size for clustered data now exist, and we encourage their wider use.[9-11] While the two articles featured in this issue help raise awareness of the challenges and some solutions to analyzing clustered data, the skills required for optimal analysis of clustered data often are beyond those of most clinician-investigators. Studies involving clustered data would greatly benefit from the expertise provided by statisticians versed in the analysis of clustered data. Several recent textbooks[3,9,12-14] and Web sites[15-17] provide good introduction to the area with realistic health care examples. Finally, the recent CONSORT statement delineating guidelines for reporting of randomized controlled trials has now been extended to the special case of cluster-randomized trials.[18]

To read or post commentaries in response to this article, see it online at http://www.annfammed.org/cgi/content/full/2/3/199.

Key words: Clustered data analysis; sample size estimation; intraclass correlation; analysis bias

Submitted March 26, 2004; accepted April 9, 2004.

References

1. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomization. *BMJ.* 1998;316:1455-1460.
2. Campbell MK, Grimshaw JM. Cluster randomized trials: time for improvement. The implications of adopting a cluster design are still largely being ignored [editorial]. *BMJ.* 1998;317:1171-1172.
3. Hox J. *Multilevel Analysis: Techniques and Application.* Mahwah, NJ: Lawrence Erlbaum; 2002.
4. Varnell SP, Murray DM, Janega JB, Blitstein MS. Design and analysis of group-randomized trials: a review of recent practices. *Am J Pub Health.* 2004; 94:393-399.
5. Localio AR, Berlin JA, Ten TR, Kimmel SE. Adjustments for center in multi-center studies: an overview. *Ann Intern Med.* 2000;135:112-123.
6. Reed JF. Adjusted chi-square statistics: application to clustered binary data in primary care. *Ann Fam Med.* 2004;2:201-203.
7. Killip S, Mahfoud Z, Pearce K. What is an intraclass correlation coefficient? *Ann Fam Med.* 2004;2:204-208.
8. Kraemer HC, Wilson T, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry.* 2002;59:877-883.
9. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Application and Data Analysis Methods.* Thousand Oaks, Calif: Sage Publications; 2001.
10. Singer J. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *J Educ Behav Stat.* 1998;24:323-355.
11. Donner A. Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold; 2000.
12. Murray DM. *Design and Analysis of Group Randomization Trials.* New York, NY: Oxford University Press; 1998.
13. Kreft I, De Leeuw J. *Introducing Multilevel Modeling.* Thousand Oaks, Calif: Sage Publications; 1998.
14. Leyland AH, Goldstein H. *Multilevel Modeling of Health Statistics.* John Wiley & Sons: London; 2001.
15. Hedeker D. Multilevel data analysis. Available at: http://tigger.uic.edu/~hedeker/ml.html.
16. Cluster for multilevel modeling. Available at: http://multilevel.ioe.ac.uk/index.html.
17. Multilevel modeling resources. University of California Los Angeles Web site. Available at http://www.ats.ucla.edu/stat/mlm/default.htm.
18. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomized trials. *BMJ.* 2004;328:702-708.