Adjusted Chi-Square Statistics: Application to Clustered Binary Data in Primary Care

James F. Reed III, PhD

St. Luke's Hospital and Health Network, Bethlehem, Pa

MORE ONLINE
www.annfammed.org

ABSTRACT

The frequency of randomized cluster trials is increasing in primary care research. These trials are differentiated by the randomization method, in which a group of individuals is randomly assigned to an intervention as a cluster rather than as individuals. Characteristically, individuals within a cluster tend to be more alike than individuals selected at random. For instance, evaluating the effect of an intervention across medical care providers at an institutional level or at a physician group practice level fits the randomized cluster model. Three examples in this article show how failure to account for the dependence introduced by unit of randomization can affect the analysis of binary data and the conclusions of randomized cluster trials. Greater consideration of the nested nature of patient, physician, and practice data would increase the quality of primary care research.

Ann Fam Med 2004;2:201-203. DOI: 10.1370/afm.41.

INTRODUCTION

Individuals in the same cluster tend to behave more alike than individuals who belong to different clusters. This dependence between individual observations is the intraclass correlation (ICC). The ICC represents the degree to which individuals from the same cluster are similar to one another compared with individuals from different clusters. Even a relatively small ICC decreases the amount of information about the effect of an intervention. If this ICC is present and positive, parameter estimates from models not accounting for this correlation might have errors that are significantly underestimated, resulting in *P* values that are too small.[1-4] Analytical methods that ignore the ICC have a tendency to underestimate the standard error of a treatment difference, thus yielding biased *P* values.[1-4] The important distinction in a randomized cluster trial is that the analysis must account for the variance introduced by the ICC. The ICC is easily estimated.[5,6]

Analytic methods specific to randomized cluster trials include estimating the ICC and computing adjustments to the Pearson chi-square as proposed by Brier[7] (χ^2_b), Rosner and Milton[8] (χ^2_{rm}), and Donner and Donald[9] (χ^2_{dd}). In addition, Rao and Scott[10] proposed 2 statistics that adjust for the clustering design effect (χ^2_{rs} and χ^2_{prs}). Details of each of these methods are included in the Appendix 1 (which is available online as supplemental data at http://www.annfammed.org/cgi/content/full/2/3/201/DC1).

The purpose of this article is to provide an introduction to the problem of analyzing binary data from clustered data in clinical research. A FORTRAN program in a executable format that produces the statistics outlined in this article is available from the author on request.*

Conflicts of interest: none reported

CORRESPONDING AUTHOR

James F. Reed III, PhD
Research Institute
St. Luke's Hospital and Health Network
801 Ostrum Street
Bethlehem, PA 18015
ReedJ@slhn.org

*An executable program that produces cluster-specific statistics χ^2_b, χ^2_{dd}, χ^2_{rm}, χ^2_{rs}, and χ^2_{prs} and sample data are available from the author (e-mail only). The input file is in free-format form (treatment, group, outcome), where the treatment variable is either a 1 or 2, the group variable identifies the cluster number (1, 2, . . ., k), and the outcome variable is binomial (1 = success, 0 = failure). Data must be in an integer format.

CLUSTER RANDOMIZED TRIALS: 3 EXAMPLES

The Diabetes Quality Improvement Project (DQIP) was sponsored by a coalition of public and private entities (Health Care Financing Administration, the American Diabetes Association Foundation for Accountability, and the National Center for Quality Assurance) and was later joined by the American Academy of Family Physicians, American College of Physicians, and the Veterans Administration. DQIP was charged to evaluate and recommend a set of diabetes-specific measures in which health care plans, physicians, clinics, and other health care providers could be compared for the purposes of accountability. These measures use a sample of diabetic patients to evaluate the performance of health plans, physician practices, or individual physicians on a consensus set of measures representing diabetes care. The DQIP measure set serves as quality-of-care indicators and a tool to be used to evaluate the performance between plans and providers for a population of patients. DQIP measures are process-oriented measures and intermediate outcome measures. The set of measures are the percentage of patients receiving 1 or more glycohemoglobin (HbA_{1c}) tests per year, percentage of patients with the highest risk HbA_{1c} level (eg, percentage of patients with HbA_{1c} >9.5%), percentage of patients assessed for nephropathy, percentage of patients receiving a lipid profile once in 2 years, percentage of patients with a low-density lipoprotein level (LDL <130 mg/dL), percentage of patients with blood pressure of less than 140/90 mmHg, the percentage of patients receiving a dilated eye examination, and the proportion of patients receiving a documented foot examination. Collectively this set of measures provides a comprehensive picture of the clinical management of patients with diabetes mellitus.

Example 1

The first example uses hypothetical data to compare the effectiveness of an intervention targeting physician care of their diabetic patients. Suppose that 17 primary care providers were provided baseline information and agreed to participate in an intervention designed to improve the care delivered to patients with diabetes. Twelve primary care providers were also provided baseline information and chose not to participate in the intervention program. The question to be assessed is the effectiveness of the intervention program. In this example, the physician is the unit of analysis, and the patient is considered nested within physicians. For each physician the number of patients that had a documented foot examination (numerator), and the total number of patient records abstracted (denominator) were recorded as follows:

For the intervention group: {30/30, 22/22, 19/19, 24/30, 28/30, 26/30, 29/30, 28/30, 27/30, 29/30, 24/30, 21/30, 14/22, 16/22, 22/32, 27/31, 16/20}.

For the nonintervention group: {14/16, 8/11, 29/35, 9/10, 6/9, 8/11, 4/6, 7/12, 7/25, 4/25, 4/23, 3/24}.

Using the Pearson chi-square, the test statistic and P value ($\chi^2 = 99.58$, $P = .0001$) indicate a significant difference between intervention group and nonintervention group compliance rates. The conclusion would be that the intervention was effective. The estimated ICC (r) is 0.2051, however. Statistics that adjust the Pearson's chi-square ($\chi^2_b = 16.17$, $P = .0001$; $\chi^2_{dd} = 17.92$, $P = .0001$; and $\chi^2_{rm} = 16.17$, $P = .0001$) all indicate a significant intervention effect. The 2 methods that adjust for the design or clustering effect ($\chi^2_{rs} = 17.13$, $P = .0001$; and $\chi^2_{prs} = 26.66$, $P = .0001$) also indicate a significant intervention effect. The effect of the large ICC (0.2051) affects the Pearson chi-square. Had the intervention effect not been as large, one would have incorrectly concluded that there was an intervention effect.

Example 2

In the second example, the same set of physicians is used to assess the proportion of patients that had an annual HbA_{1c} indicator. The number of patients that met this quality indicator (numerator) and the total number of patient records abstracted (denominator) were recorded as follows:

For the intervention group: {29/30, 22/22, 11/19, 29/30, 29/30, 27/30, 28/30, 29/30, 23/30, 26/30, 30/30, 29/30, 20/22, 20/22, 18/32, 27/31, 15/20}.

For the nonintervention group: {14/16, 10/11, 27/35, 8/10, 9/9, 7/11, 4/6, 5/12, 17/25, 23/25, 19/23, 18/24}.

Again, if one were to use the Pearson chi-square, the test statistic and P value ($\chi^2 = 11.7694$, $P = .0007$) would indicate a significant difference between the intervention group and nonintervention group compliance rates. The conclusion would be that the intervention was effective. The estimated ICC is 0.0938. Statistics that adjust the Pearson's chi-square ($\chi^2_b = 3.51$, $P = .0579$; $\chi^2_{dd} = 3.81$, $P = .0513$; and $\chi^2_{rm} = 3.51$, $P = .0579$) all indicate a nonsignificant intervention effect. The 2 methods that adjust for the design or clustering effect ($\chi^2_{rs} = 4.22$, $P = .0377$; and $\chi^2_{prs} = 1.16$, $P = .2812$) contradict one another. Cluster-specific analytic methods use the cluster as the unit of analysis and have the same consequences as many studies (eg, smaller sample sizes—number of physician practices as well as the number of patients within each physician practice—have a tendency to reduce the power of the study). In this example, a relatively small ICC can affect the analysis of randomized cluster trials.

Example 3

The third hypothetical example again uses the same set of physicians in assessing the annual lipid profile indica-

tor. The number of patients that met this quality indicator (numerator) and the total number of patient records abstracted (denominator) were recorded as follows:

For the intervention group: {21/30, 13/22, 10/19, 24/30, 13/30, 14/30, 18/30, 24/30, 17/30, 22/30, 20/30, 21/30, 14/22, 19/22, 14/32, 25/31, 14/20}

For the nonintervention group: {11/16, 7/11, 23/35, 7/10, 6/9, 7/11, 4/6, 6/12, 17/25, 20/25, 17/23, 14/24}

First look at the ICC (r = 0.0141). I would not expect that the magnitude of this ICC to influence the analysis. The Pearson chi-square (χ^2 = 0.3637, P = .5519) indicates that there is no difference between the intervention and nonintervention physician groups in the proportion of patients that have completed an annual lipid profile. When one has a randomized cluster design, however, the appropriate analysis should be reflected in the analysis. Adjustments to the Pearson chi-square (χ^2_b = 0.27, P = .6088; χ^2_{dd} = 0.28, P = .6038; and χ^2_{rm} = 0.27, P = .6088) all agree as expected. The statistics that use the design effect also concur (χ^2_{rs} = 0.35, P = .5624; and χ^2_{prs} = 0.09, P = .7528). There is no apparent intervention effect between the 2 physician groups regarding the proportion of patients that have completed an annual lipid profile.

DISCUSSION

Sometimes interventions in randomized clinical trials are allocated to groups rather than individual patients. Such randomization is cluster allocation, or cluster randomization, and is found with increasing frequency in health services research and in primary care. It appears that most of these trials may not account appropriately for the clustering in their analysis. Failure to account for the lack of independence between individual observations and the cluster to which they belong can lead to inappropriate analyses. Likewise, inappropriate analysis of cluster trials can lead to inaccurate results and misleading conclusions.[2,4] The randomized cluster design could be easily expanded to observational studies, where data are collected on patients nested within physicians or in which physicians are nested within a practice.

For some interventions, it may be necessary to randomize clusters rather than individuals. For instance, in a typical randomized cluster design, randomizing the physician rather than the patient prevents contamination of the intervention, because patient management by a physician tends to be the same from patient to patient. Standard statistical methods are not appropriate when analyzing cluster randomized trials. There is no consensus, however, as to which analytical approach should be used to analyze all cluster randomized trials.[3]

Which analytic method is preferred when analyzing binary responses from a randomized cluster trial? Adjustments to the Pearson chi-square are based on the clustering and homogeneity of design effects for the treatment groups, are computationally friendly, and provide excellent design-specific alternatives. Their behavior in relatively small numbers of clusters within a treatment group, however, may be problematic. Analysis-of-variance methods are simple and may be used in multifactorial designs. A disadvantage is that they do not stabilize the variances—a necessary requirement for analysis of variance. I prefer the adjustments to the Pearson chi-square statistic and recommend adapting the analytical plan to the study design by analyzing binomial responses from cluster trials using cluster-specific methods.

Cluster-randomized trials represent an important experimental design. They are particularly relevant when evaluating interventions at the clinic level, with physicians, or in physician group practices. Serious design and analysis implications abound, and the use of clusters as the unit of randomization must be justified. Sample sizes, the number of individuals within the cluster variable, and the number of clusters usually need to be larger, and the analysis must certainly allow for the cluster design.[1]

To read or post commentaries in response to this article, see it online at http://www.annfammed.org/cgi/content/full/2/3/201.

Key words: Clinical trials; cluster analysis; epidemiologic studies

Acknowledgments: Referees serve as a valuable resource to this journal. I am grateful to the anonymous referees for their thorough reviews, helpful comments, and suggestions that have improved the presentation of the paper.

Submitted February 12, 2003; submitted, revised, March 4, 2003; accepted March 24, 2003.

References

1. Fayers PM, Jordhoy MS, Kaasa S. Cluster-randomized trials. *Palliat Med*. 2002;16:69-70.
2. Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. *Acad Emerg Med*. 2002;9:330-341.
3. Mollison J, Simpson JA, Campbell MK, Grimshaw JM. Comparison of analytical methods for cluster randomized trials: an example from a primary care setting. *J Epid Biostat*. 2000;5:339-348.
4. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. *Fam Pract*. 2000;17:192-196.
5. Shrout PE, Fleiss JL. Intraclass correlation: Use in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
6. Bodian CA. Intraclass correlation for two-by-two tables under three sampling designs. *Biometrics*. 1994;50:183-193.
7. Brier SS. Analysis of contingency tables under cluster sampling. *Biometrika*. 1980;67:591-596.
8. Rosner B, Milton RC. Significance testing for correlated binary outcome data. *Biometrics*. 1988;44:505-512.
9. Donner A, Donald A. The statistical analysis of multiple binary measurements. *J Clin Epidem*. 1988;41:899-906.
10. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics*. 1992;48:577-585.