Program evaluation

Statistical and design issues in studies of groups

P Cummings, T D Koepsell

Accounting for within-group correlation

In the current issue of *Injury Prevention*, there are two studies of groups. Ozanne-Smith and colleagues studied the impact of an educational injury prevention program on two communities in Australia,[1] and Lindqvist *et al* conducted a similar study in Sweden.[2]

Studies of naturally occurring groups are common for several reasons. For example, Levy and colleagues, in the United States, wanted to know if differences in state rules regarding driver license renewal were associated with the crash mortality rates of elderly drivers.[3] Their study compared mortality rates in different states, because the exposure, licensing rules, applied to everyone in each state. Kannus *et al*, in Finland, wanted to study the association of hip protectors with hip fracture.[4] They feared that if they randomized elderly individuals, it might be hard to keep those assigned to the control group from using the hip protectors, and this might attenuate the study's ability to detect any difference in outcomes. Therefore, they randomized treatment centers, rather than individuals within each center; the goal was to have all patients in some centers wear hip protectors while all patients in control centers would go without these devices. Other examples of groups in studies include students in schools,[5] patients in physician practices,[6] and workers in stores.[7]

A few textbooks and many articles have commented upon design and statistical issues that apply to studies of groups.[8–15] In this commentary we will discuss one feature of group studies: individuals within groups may be more alike in their propensity to have a study outcome compared with individuals between groups.

When individuals are in groups (clusters), investigators need to consider the possibility that children in a classroom, patients in a practice, players on a baseball team, or citizens of a country may be more (or less) prone to have the outcome under study than persons in other classrooms, practices, etc. The statistical literature refers to this situation using many terms: within-cluster correlation, lack of statistical independence within groups, and between-cluster variability. Within-group correlations can come about by any of several mechanisms, including: (1) shared exposure to the same physical or social environment; (2) self selection in belonging to the group; (3) sharing of behaviors, ideas, or diseases among members of the group. The consequence of within-cluster correlation of outcomes is that a study of persons in groups may be less efficient (that is, have less statistical power) compared with a study of the same total number of individuals, but in which assignment to exposure was by individual rather than group.

Everyone can be thought of as being a member of a group. Everyone lives in a certain geographic area, has a family (at least parents), has certain beliefs or habits that others share, and so on. Membership in a group becomes an issue when the study intervention or exposure is applied to groups. A study of state laws, for example, is necessarily a study of groups.

Imagine that we wish to prevent falls by providing individuals with a new shoe made by Acme Sports, Inc. We plan to randomize individuals to the new footwear or a control arm in which people wear their usual shoes. We estimate that in the control arm, 50% of study subjects will have a fall within six months of the start of the trial. Due to cost constraints, we plan on a study with 1000 members in each trial arm; we calculate that this will give us 90% power to detect a decrease in the proportion of subjects with a fall to 42.67%, using $p<0.05$ as our criteria for a statistically significant result.

Acme Sports will finance our study, but they are concerned about the cost. They suggest that we can save money by recruiting people through churches, randomizing the churches, and assigning the members of each church to one trial arm. Local churches agree that we can pass out Acme shoes with the collection plates. There are 20 churches in the study area, each with 100 adult members, and they all agree to participate, guaranteeing us the needed 2000 study subjects. But we wonder how this might affect study power. People within churches not only share similar religious beliefs, but they may be more alike in regard to income, politics, health habits, and, perhaps, their propensity to fall. Concerned that randomizing 20 churches might not be the same as randomizing 2000 individuals, we might turn to a statistician for help.

Table 1 Sample size calculations for the hypothetical Acme shoe study. It is assumed that the randomized groups all include 100 study subjects. Statistical significance set at $p<0.05$ and power set at 0.9. 50% of controls will fall; we wish to detect an invention effect that would reduce the per cent falling in the intervention arm to 42.67%. Study subjects are to be evenly divided between intervention and control arms

Values of ρ*	No of groups	No of subjects
0.000	20	1996
0.001	22	2194
0.005	30	2986
0.01	40	3974
0.05	119	11878
0.1	218	21758

*ρ is the intraclass correlation coefficient.

The statistician will tell us that the loss of study power will depend on the intraclass correlation coefficient, often designated by the Greek letter ρ (rho). The intraclass correlation coefficient is a measure of within-group correlation of outcomes. If ρ is zero, then group randomization can result in relatively little loss of study power. If ρ is 1, this means that within each group all members will have the same outcome; as a consequence, it would be as if our study sample were reduced to just 20 members. Values of ρ tend to be larger in small groups, such as a family, and smaller in large groups, such as a state, because the degree of clustering often depends upon the interaction of group members; family members are usually more alike than persons in different areas of a large geographic region. Unfortunately, the influence of ρ on study power is directly related to group size. Studies with a few large groups are often very inefficient. For our Acme shoe study, the statistician prepared a table of sample sizes for various possible values of ρ (table 1). On reviewing this table, our enthusiasm for randomization by group may be somewhat dampened. An intraclass correlation coefficient of only 0.01 would require that we double the size of our study population, in order to have the same power as a study that randomized individuals.

There are lessons in table 1 for investigators. If one can study individuals,

rather than groups, without much additional cost or effort, take that course. Secondly, studying more groups always increases power, even if this means studying fewer people per group to keep constant the total number of individuals studied. Third, if groups are to be studied prospectively, do power calculations that account for the grouped nature of the data. Estimates of ρ for study planning purposes may sometimes be obtained from previous studies or from analyses of pre-existing data from the anticipated study setting.[9]

Investigators should use analytic methods that account for within-group correlations. Journals are lax about publishing grouped trials analyzed by methods suitable for studies of individuals.[16 17] For example, the studies by Kannus *et al*[4] and Wassell *et al*,[7] mentioned earlier, were analyzed without accounting for the grouped nature of the data. This practice can be misleading, as estimates of precision (such as confidence intervals), may be too narrow and p values may be unjustifiably small.

There are several statistical methods available for the study of grouped data. Some statistical packages offer a variety of methods for this type of analysis; the choices can be a bit overwhelming and not all methods produce the same result.[8 10–13] To make a choice, investigators should read the literature on these methods and possibly seek advice from someone with experience in this area.

STUDIES OF TWO GROUPS

The studies from Australia[1] and Sweden,[2] in this issue of *Injury Prevention*, failed to account for the grouped nature of their interventions. They could not measure and account for any within-group correlation of outcomes because they only had two groups. Imagine that you wished to compare the weights of men and women and you sampled, at random, one man and one woman from a population. The man weighed 90 kg and the woman weighed 45 kg. These weights are different (by 45 kg), but are they statistically different? In other words, could the difference we found be due to chance? Unfortunately, no statistical comparison is possible. To make a comparison, you need some estimate of the variation in weight of the men and the women in the population. But since you have only one of each, no estimate of variation in weight by sex is possible. Therefore, it is not possible to estimate a p value or a confidence interval for a comparison of the weights. This analogy may help explain why a study of two groups cannot estimate within and between group correlations; there is only one group in each arm of the study. A statistical comparison of individuals in the groups can be made, but without any estimate of the within-group correlation, the p values and confidence intervals from such a comparison may be too small. In order to estimate within-group correlation of outcomes in each exposure arm of a study, at least two groups are needed per arm. In practice, more groups are desirable.

Does this mean that studies of two groups are not worthwhile? We do not take that position. Hundreds of studies with just two groups have been done. If a difference in outcomes is large and not plausibly explained by other factors, such a study can be persuasive. Similarly, the study may be useful if the difference in outcomes is zero; we don't need a p value to tell us that there was no difference, although we might still desire valid confidence intervals, to assess what degree of difference might be compatible with the data.[18]

However, because of concern that observations within groups may not be fully independent, we suggest that readers should be aware that confidence intervals and p values may be too small in some studies of two groups. Investigators should also be aware of this, and mention this as a limitation of their study. If there is a reasonable external estimate for the intraclass correlation, the investigators could test the sensitivity of key results to clustering by adjusting their confidence intervals or p values for this degree of within-group correlation. If there is no empirical estimate, the intraclass correlation coefficient might still be varied across a plausible range of values to evaluate the robustness of conclusions. Finally, investigators should be aware that if they are planning a study of groups, they might be better off studying several groups rather than just two. Use of power calculations that account for within-group correlation can help them make sensible decisions. Two recently published textbooks provide an excellent introduction to studies of groups, including information about power calculations and analytic methods.[8 10]

Injury Prevention 2002;**8**:6–7

Authors' affiliations
P Cummings, T D Koepsell, Harborview Injury Prevention and Research Center and the Department of Epidemiology, University of Washington, Seattle, Washington, USA

Correspondence to: Dr Peter Cummings, Harborview Injury Prevention and Research Center, 325 Ninth Avenue, Box 359960, Seattle, WA 98104–2499, USA; peterc@u.washington.edu

REFERENCES

1 **Ozanne-Smith J**, Day L, Stathakis V, *et al*. Controlled evaluation of a community based injury prevention program in Australia. *Inj Prev* 2002;**8**:18–22.
2 **Lindqvist K**, Timpka T, Schlep L, *et al*. Evaluation of a child safety program based on the WHO Safe Community model. *Inj Prev* 2002;**8**:23–6.
3 **Levy DT**, Vernick JS, Howard KA. Relationships between driver's license renewal policies and fatal crashes involving drivers 70 years or older. *JAMA* 1995;**274**:1026–30.
4 **Kannus P**, Parkkari J, Niemi S, *et al*. Prevention of hip fracture in elderly people with use of a hip protector. *N Engl J Med* 2000;**343**:1506–13.
5 **Grossman DC**, Neckerman HJ, Koepsell TD, *et al*. Effectiveness of a violence prevention curriculum among children in elementary school: a randomized controlled trial. *JAMA* 1997;**277**:1605–11.
6 **Grossman DC**, Cummings P, Koepsell TD, *et al*. Firearm safety counseling in primary care pediatrics: a randomized, controlled trial. *Pediatrics* 2000;**106**:22–6.
7 **Wassell JT**, Gardner LI, Landsittel DP, *et al*. A prospective study of back belts for prevention of back pain and injury. *JAMA* 2000;**284**:2727–32.
8 **Murray DM**. *Design and analysis of group-randomized trials*. New York: Oxford University Press, 1998.
9 **Koepsell TD**. Epidemiologic issues in the design of community intervention trials. In: Brownson RC, Pettiti D, eds. *Applied epidemiology: theory to practice*. New York: Oxford University Press, 1998: 177–211.
10 **Donner A**, Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000.
11 **Campbell MJ**. Cluster randomized trials in general (family) practice research. *Stat Methods Med Res* 2000;**9**:81–94.
12 **Feng Z**, Diehr P, Peterson A, *et al*. Selected statistical issues in group randomized trials. *Annu Rev Public Health* 2001;**22**:167–87.
13 **Neuhaus JM**. Assessing change with longitudinal and clustered binary data. *Annu Rev Public Health* 2001;**22**:115–28.
14 **Neuhaus JM**, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998;**54**:638–45.
15 **Localio AR**, Berlin JA, Ten Have TR, *et al*. Adjustments for center in multicenter studies: an overview. *Ann Intern Med* 2001;**135**:112–23.
16 **Donner A**, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *Int J Epidemiol* 1990;**19**:795–800.
17 **Simpson JM**, Klar N, Donner A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health* 1995;**85**:1378–83.
18 **Altman DG**, Bland MJ. Absence of evidence is not evidence of absence. *BMJ* 1995;**311**:485.