Published in final edited form as:

Appl Nurs Res. 2007 August ; 20(3): 152–157.

Sample size considerations when groups are the appropriate unit of analyses

Georgia Robins Sadler, Ph.D., MBA, BSN,

Clinical Professor of Surgery, School of Medicine, University of CA, San Diego, Associate Director of Outreach, Rebecca and John Moores UCSD Cancer Center, 3855 Health Sciences Drive # 0850, La Jolla, CA 92093-0850, 858-534-7611. gsadler@ucsd.edu

Celine Marie Ko, M.A., M.S.,

Research Associate, Rebecca and John Moores UCSD Cancer Center, 3855 Health Sciences Drive # 0850, La Jolla, CA 92093-0850, 858-534-7611. cmko@ucsd.edu

Jennifer Alisangco, BS, RN,

Intern, Minority Training Program in Cancer Control Research, Rebecca and John Moores UCSD Cancer Center, 3855 Health Sciences Drive # 0850, La Jolla, CA 92093-0850, 858-534-7611. jencinitas@sbcglobal.net

Bradley P. Rosbrook,

Senior Statistician, Rebecca and John Moores UCSD Cancer Center, 5855 Copley Drive #0905, La Jolla, CA 92093-0905, 858-300-1062. brosbrook@ucsd.edu

Eric Miller, B.S., and

Student, Rebecca and John Moores UCSD Cancer Center, 3855 Health Sciences Drive # 0850, La Jolla, CA 92093-0850, 858-534-7611. We do not have this student's email any longer

Judith Fullerton, Ph.D., CNM, FACNM

Senior Technical Officer for Monitoring & Evaluation, Project Concern International, 5151 Murphy Canyon Road, San Diego, CA 92123, 858-693-7585. jfullerton@san.rr.com

Abstract

This paper discusses issues to be considered by nurse researchers when groups should be used as a unit of randomization. Advantages and disadvantages are presented, with statistical calculations needed to determine effective sample size. Examples of these concepts are presented using data from the *Black Cosmetologists Promoting Health Program*. Different hypothetical scenarios and their impact on sample size are presented. Given the complexity of calculating sample size when using groups as a unit of randomization, it's advantageous for researchers to work closely with statisticians when designing and implementing studies that anticipate the use of groups as the unit of randomization.

Keywords

MeSH; Intraclass correlation coefficient; Group randomization; Recruitment; Effective Sample Size; Nursing Research

INTRODUCTION

In most research studies, the unit of randomization is the individual. There are clinical study designs, however, where the appropriate unit of randomization will be groups of individuals rather than the individuals themselves. In such instances, groups (clusters) will be knowingly selected as the unit of randomization. When this is anticipated, the study design, sample size, and statistical analysis will need to be adjusted accordingly. Special statistical considerations must be taken into account to arrive at both the correct sample size and the optimal methods of data analysis. The degree to which the individual's membership in the cluster influences the participant's individual responses must also be evaluated before the data can be interpreted and conclusions drawn.

In other instances, the unit of randomization is intended to be the individual, but because the individuals will be drawn from existing groups or will share important common bonds, the concern arises that they may not function independently of each other. When this occurs preliminary statistical evaluations should be done on pilot data and again on the full sample to evaluate the amount of inter-dependency within the sample to determine whether the individuals can be considered to act independently of each other (Lake, Kammann, Klar, & Betensky, 2002).

The reverse is also possible. The study design could intentionally select groups as the unit of randomization. However, preliminary statistical analysis of pilot data may show that for the purposes of the study, the members of the groups function sufficiently independently of each other that the individual members within the group can be treated as the unit of randomization. This situation reduces the number of individuals required for the sample by allowing individuals as the unit of randomization and permits standard methods of data analysis to be followed.

This paper discusses the advantages and disadvantages of using groups as the unit of randomization. It also provides information and resources for nurse researchers to assist them to understand the statistical procedure for determining the degree of inter-dependence among individuals within the sample and the statistical steps needed to derive the required sample size when a high level of intra-group dependence mandates using groups as the unit of randomization (Donner & Klar 1996; Wears 2002).

Pros and Cons of Using Groups as the Unit of Randomization

There are many advantages to the use of groups as the unit of randomization. The first advantage is the reduction of contamination among study subjects. Often it is impossible to assure that individuals assigned to one arm of a study do not influence individuals assigned to the other arm. For example, in a school-based dietary intervention program, randomizing all children from a single school into an experimental and control arm does not adequately prevent contamination. Children imitate each other and their parents may talk among themselves. This problem becomes easy to minimize by using schools as the unit of randomization.

Second, ethical considerations may best be served when the unit of randomization is a group. Ethical concerns can arise when one study group experiences significant benefits over the other study group. For example, a program to test the value of providing three meals to underprivileged school children at the school site versus no meals would be ethically and morally difficult to implement by randomizing children at the same school site into the two study arms. Using children at separate school sites would reduce these ethical concerns, while allowing an important hypothesis to be tested.

The third advantage to randomization by group is that the group may offer the easiest or least expensive way to introduce an intervention. Indeed in some instances, randomization by individuals may not even be possible. Studies of the transmission of ideas, for example, recognize the role of group innovators and their followers.

Fourth, the group may intentionally be used as the unit of randomization because its role in the intervention is an issue that is being evaluated. For example, can a greater degree of behavior modification be achieved when women receive information in a church or social group setting where there is the opportunity to reinforce each other than when such information is delivered one-on-one? When people interact with others involved in the same learning process, peer influence may make them more likely to adhere to the recommendations. For example, couples who were taught in an interactive group setting were more likely to have improved physical fitness, decreased saturated fat intake, and decreased levels of low-density lipoprotein cholesterol compared to those couples who received mailed education materials (Burke, Giangiulio, Gilliam, Beilin, & Houghton, 2003). Thus, a researcher might wish to intentionally design an intervention program directed towards interactive groups rather than one directed to individuals to more efficiently promote lifestyle improvement.

Fifth, study designs that administer interventions and/or evaluations to groups are usually more economical than interventions that must work with each member of the sample individually. Further, the delivery of the intervention or evaluation will likely be more consistent when it can be offered to many individuals simultaneously and irregularities in the intervention's or evaluation's delivery can be more readily identified.

In spite of the advantages, there are several disadvantages encountered when using groups as the unit of randomization. The similarities and interdependence of the individuals within a group cause the data to be correlated, deflates variances, and shrinks the significance of any one response (Donald & Donner 1987). The result, when dealing with groups, is a loss in power to detect the real intervention effect. To make up for this loss of power, either more units of randomization (groups) or more individuals per group must be added to the study's sample.

Furthermore, one may have to adjust for an imbalance between treatment groups on important individual subject characteristics, such as smoking behavior, exercise levels, or sociodemographic characteristics (Cosby, Howard, Kaczorowski, Willian, & Sellor, 2003). When a sufficient number of groups are being studied, a way to lessen this risk is to pre-stratify the clusters according to common factors (Lewsey, 2004), e.g., stratify by smokers vs. non-smokers. However, even this solution must be carefully evaluated because it can introduce substantial new variance among each of the groups or clusters. For example, an analysis of the sample of smokers and non-smokers, may disclose that the average age of the smokers is significantly higher than the average age of the sample of non-smokers. Doing an analysis of the sample characteristics part way through the study enables the researcher to identify this disparity in the two sub-samples and make strategic changes in the remaining sample recruitment.

Finally, there is the problem of finding enough like groups in the community or region to create a large enough sample of groups to yield statistically reliable conclusions (Connelly, 2003). Residents from a geographic location remote from the primary study site may be ideal subjects for the non-intervention, control arm of the study both from a convenience standpoint and because contamination from the experimental arm has been minimized. However, going beyond the region to secure additional groups usually increases the cost, administrative complexity of conducting the study, and problems of maintaining quality control over all components of the study.

Determining Required and Optimal Sample Size when the Groups are the Unit of Randomization

Step 1: Calculate the required sample size needed without clusters The first step in determining the appropriate sample size for a study that intends to use groups as the unit of analysis is to calculate the size of the sample that would be needed if no clustering were present. To produce statistically significant conclusions when data are clustered or pulled from groups, sample size may have to be increased. In the worst case where clustering is maximum, the numbers of clusters or groups may have to be increased until the number of clusters is the same as the size required for a simple random sample of individuals.

Sample size calculations are relatively straightforward when calculating the size required for a simple random sample of individuals. This calculation is done by estimating the required number of individuals based on the desired power (typically .7 to .9), significance level desired (typically .01 or .05), and anticipated detectable difference between means (effect size).

Step 2: Determine the intraclass correlation coefficient (ICC) In contrast, when data are clustered, finding the sample size required to produce statistically significant results involves more sophisticated calculations (Hayes & Bennett 1999; Turner, Prevost, & Thompson, 2004; Campbell, Thomson, Ramsay, MacLennan, & Grimshaw, 2004). Determining the optimal sample size where groups or clusters are involved requires calculations involving the number of groups or clusters, number of individuals in the groups, in addition to the power, significance level, and effect size the investigator seeks to achieve (Donner, Piaggio, & Villar, 2003). Table 1 offers a glossary of these statistical terms. The investigator also needs to include in this calculation an estimate of the interdependence expected among the individuals within each cluster or the intraclass correlation coefficient (ICC or ρ). The ICC measures the relatedness of clustered data by comparing variance within clusters with variance between clusters. The ICC can either be selected from previous work with similar groups or calculated from pilot study data.

Data collected in a pilot study enables the research team to calculate the degree of independence versus interdependence among the members within each group. Specifically, the researcher examines the degree to which subjects' presence in one study arm predicts their answers or personal characteristics when compared to other members of that same study arm. The higher the likelihood of prediction, the higher the ICC, and the higher the number of clusters needed. The pilot study's ICC can then be included in the calculation of the definitive study's effective sample size.

The intraclass correlation coefficient assesses the degree of variation among individuals within a group. The ICC ranges from 0 to 1; zero indicates that there is no relationship among responses within a cluster and one indicates that all responses within a cluster are the same. The preferred ICC values in large human studies consisting of hundreds of subjects are typically less than 0.01 - 0.02 (Killip, Mahfoud, & Pearce, 2004). The acceptable value does depend on the sample size, with smaller studies having a higher acceptable ICC than studies with larger sample sizes. The number of ICC calculations that a researcher should conduct before deciding whether clustering is or is not present depends entirely on the level of assurance the investigator seeks. This in turn depends on the critical importance of the research questions being asked.

Step 3: Calculate the effective sample size Having estimated the ICC from the pilot study data, the next step is to determine the sample size that will be needed for the definitive study after correcting for clustering: the effective sample size (ESS). The ESS can be calculated with the equation: mk/DE (number of groups [m] multiplied by number of individuals [k] divided by the design effect [DE]). The design effect is a correction factor that incorporates the ICC, or presence of clusters. A further explanation of design effect is offered in Table 1.

The degree of interdependence of the individuals within each group (ICC) will determine if the effective sample size is the same as, or smaller than, the sample size that was originally estimated to be adequate if no clustering were present. When the effective sample size is smaller than the required sample size without clusters, the sample size for the definitive study must be increased to produce statistically valid conclusions. Statistical software, such as SPSS (www.spss.com) and SAS (www.sas.com) offer programs that can be used to calculate the measures of group correlation and sample size.

Considerations for Sample Size Adjustment There are at least three points in a study's design when the investigator must decide whether to plan the study's data analysis as if the unit of randomization is the group or cluster rather than the individual, and to use the different statistical treatments required by clustered data: 1) while designing the study methodology in light of the pilot data or the results of similar investigations; 2) at the midpoint of data collection for the full study; and 3) prior to data analysis.

Even when a pilot study has been undertaken and found to have a low intraclass correlation coefficient (meaning, therefore, that the sample can be calculated as if individuals are the unit of randomization), the researcher should again evaluate the ICC near to or at the midpoint of data collection for the definitive sample. The clustering may not be apparent in the small pilot study, but may become apparent when the ICC is recalculated.

Finding a medium to high intraclass coefficient (0.5 to 1) will alert the researcher to the possible need for adjustments to the sample (number of individuals or number of groups) that needs to be recruited. When the ICC equals one, the individuals in each group are so dependent upon each other that they are likely to respond identically to any intervention. As a result, each group must be considered as if it were a distinct unit of randomization making the effective sample size equivalent to the number of groups in the sample. Hence the number of groups must be increased to equal the sample size required under individual sampling without clusters (as calculated in step 1 above). On the other hand, if the computed value equals zero, this shows there is no correlation among the responses of the individuals in each group. Each member of the group can now be considered as a distinct individual making the effective sample size the number of groups times the number of individuals in each group. Standard methods of data analysis can now be used. Rarely, however, is the intraclass correlation coefficient likely to be at the extremes of zero or one. Therefore, the investigator must decide at what ICC score the unit of randomization will be considered the group and when it will be considered the individual participant.

When the possibility of clustering is raised only after data collection for the study has been completed, the computation of the ICC (steps 1 through 3 above) should be followed before data analysis is initiated to determine if the effective sample size was, in fact, achieved. However, if done only prior to data analysis, there is no guarantee that the study will have enough power to detect a statistically significant difference. If it can be documented that the effect of clustering was, in fact, minimal, then data analysis can proceed as originally planned. Alternatively, if a strong cluster effect is detected, then the conclusions drawn from the study may be limited.

Application of the Concepts

The examples cited in the following sections of this article are derived from the *Black Cosmetologist Promoting Health Program*, a study for which the primary author served as Principal Investigator (PI). This study will be used to illustrate the use of groups as units of randomization (Sadler, Ko, Escobar, White, Neal, Lee et al, 2005; Sadler, Ko, Butcher, Lee, Neal et al. 2004; Sadler, Thomas, Gebrikristos, Dhanjal, Mugo et al. 2000; Sadler, Thomas, Dhanjal, & Wright, 1998). From 1998-2000, women were recruited from 21 hair dressing

salons in a Southern California community to participate in this randomized controlled clinical trial that tested the efficacy of a cancer education intervention. First, cosmetologists from different salons were approached about participating in the research study. The stylists encouraged their clients to sign a consent form to participate in the study and to complete a baseline questionnaire regarding their cancer and diabetes health beliefs, knowledge, and screening practices. Research assistants also helped the cosmetologists to recruit women to participate in the study by approaching them while they were waiting for their appointments at the beauty salons. In this study, salons are the unit of randomization, and women clients of each salon are the members of each unit or cluster. Once the full complement of subjects was recruited, the salons were randomized to either the control arm, which delivered diabetes education, or the experimental arm, which delivered breast cancer education.

The goal was to recruit approximately 50 women per salon. The number of women recruited per salon ranged from 17 to 88 women. The mean number of women per salon was 50 women, with a standard deviation of 16.83 women. Table 2 demonstrates a hypothetical situation in which any of three variables are manipulated (Killip et al. 2004). The first option depicts the case where the number of salons recruited into the study remains constant ($k=21$). Increasing the number of women recruited per salon results in an increase in effective sample size (ESS), which also results in an increase in power. As shown in this table, a researcher working with a set number of recruitment sites must ensure that a sufficient number of individuals are recruited from each site to obtain an effective sample size.

In the second case the number of women per salon remains constant ($m=50$). Increasing the number of salons while holding the number of women recruited from each salon constant leads to an increase in both the ESS and power. Consider the situation where the principal investigator originally designed a study to recruit 10 salons, and a target of 50 women per salon. Upon preliminary analysis, if the PI discovered high intraclass correlation coefficients in the outcome variables, the PI might want to recruit more salons to increase ESS and power.

The third case demonstrates that even a large sample size may not equate to sufficient power if one does not take into account the ICC. For instance, if all 1050 women were recruited from the same salon, the ESS = 56. In essence, there is a sample size of only 56 individuals, which results in a low power (.411). However, as in this study, even when ICC = .025, a total sample size of 1050 women still leads to an ESS equal to 472 and higher power (.582). These calculations were done in anticipation of using a 2-sample t-test to compare means between the intervention and control group, with a small effect size, and an alpha of .05.

Table 3 demonstrates the effect of ICC on the effective sample size and power for the *Black Cosmetologist* study. As the intraclass correlation increases, effective sample size and study power both decrease.

CONCLUSION

The use of groups as the unit of randomization is a rapidly evolving methodologic consideration. There are many areas of uncertainty, and research is continuing within the field. It is advantageous for nurse researchers to work closely with their statistician when designing studies that anticipate the use of groups, rather than individuals, as the unit of randomization. Study investigators who are aware of the problems that may arise when data are clustered and anticipate the possibility that adjustments might need to be made in sample size or the approach to data analysis will avoid unnecessary limitations to the interpretation and generalizations of their study results.

ACKNOWLEDGEMENTS

We would like to acknowledge National Cancer Institute (NCI) Grant # R25 CA 65745 and Cancer Center Core Grant #5 P30 CA023100-22, the National Institutes of Health, Division of National Center on Minority and Health Disparities EXPORT Grant # P60 MD 00220, the NCI Minority Institution/Cancer Center Partnership Program Grants # U56 CA 92079 and #U56 CA 92081, NCI Increasing Diversity in Cancer Control Research Grant # R25 CA 07583. We also wish to acknowledge Karen Clarke for her valuable statistical input on the paper.

REFERENCES

- Burke V, Giangiulio N, Gilliam HF, Beilin LJ, Houghton S. Physical activity and nutrition programs for couples: A randomized controlled trial. *Journal of Clinical Epidemiology* 2003;56:421–432. [PubMed: 12812815]
- Campbell MK, Thomson S, Ramsay CR, MacLennan GS, Grimshaw JM. Sample size calculator for cluster randomized trials. *Computers in Biology and Medicine* 2004;34(2):113–25. [PubMed: 14972631]
- Connelly LB. Balancing the number and size of sites: an economic approach to the optimal design of cluster samples. *Controlled Clinical Trials* 2003;24(5):544–559. [PubMed: 14500052]
- Cosby RH, Howard M, Kaczorowski J, Willan AR, Sellor JW. Randomizing patients by family practice: Sample size estimation, intracluster correlation and data analysis. *Family Practice* 2003;20(1):77–82. [PubMed: 12509376]
- Donald A, Donner A. Analysis of data arising from a stratified design with the cluster as unit of randomization. *Statistics in Medicine* 1987;6(1):43–52. [PubMed: 3576016]
- Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology* 1996;49(4):435–439. [PubMed: 8621994]
- Donner A, Piaggio G, Villar J. Meta-analyses of cluster randomization trials: Power considerations. *Evaluation & the Health Professions* 2003;26(3):340–51.
- Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* 1999;28(2):319–326. [PubMed: 10342698]
- Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine* 2004;2(3):204–208. [PubMed: 15209195]
- Lake S, Kammann E, Klar N, Betensky R. Sample size re-estimation in cluster randomization trials. *Statistics in Medicine* 2002;21(10):1337–50. [PubMed: 12185888]
- Lewsey JD. Comparing completely and stratified randomized designs in cluster randomized trials when the stratifying factor is cluster size: A simulation study. *Statistics in Medicine* 2004;23(6):897–905. [PubMed: 15027079]
- Sadler GR, Ko CM, Escobar R, White M, Neal T, Lee S, et al. African American women's perceptions of their most serious health problems. *Journal of the National Medical Association* 2005;97(1):31–40. [PubMed: 15719869]
- Sadler GR, Meyer M, Ko CM, Butcher C, Lee S, Neal T, et al. Diabetes knowledge among African American women. *Diabetes Educator* 2004;30(4):676–685. [PubMed: 15669783]
- Sadler GR, Thomas AG, Gebrikristos B, Dhanjal S, Mugo J. Black cosmetologists promoting health program: Pilot study outcomes. *Journal of Cancer Education* 2000;15(1):33–37. [PubMed: 10730801]
- Sadler GR, Thomas AG, Dhanjal S, Wright F. Breast cancer screening adherence in African American women: Black cosmetologists' promoting health program. *Cancer Supplement* 1998;83(8):1836–1869.
- Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine* 2004;23(8):1195–1214. [PubMed: 15083478]
- Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. *Academic Emergency Medicine* 2002;9(4):330–341. [PubMed: 11927463]

Table 1

Glossary of Terms

	Range	Desired direction of range	Definition
Power ($1 - \beta$)	0 - 1	Closer to 1 is better	The probability of identifying a pre-determined meaningful difference between groups. Power increases with sample size.
Significance level (α or p-value)	0 - 1	Low is better Less than 0.01 or 0.05. Typically 0.05 is used in social science studies.	The degree to which an observed difference between sample means is a reliable one. The value is a predetermined probability level used to assess the null hypothesis (there are no differences between the means). Usually, the larger the sample size, the greater the power of a study to detect a significant difference.
Effect Size (Cohen's d)	Range: 0 - 1 if you are looking at percentages, but any number if you are looking at means.	Depends on the desired strength of the relationship between the independent variable and the outcome variable.	The degree to which two experimental do not overlap with one another. It is a measurable difference that reveals whether the manipulation of an independent variable truly separated the values of populations from one another. Effect size influences power. The greater the effect size, the greater the power of the study.
Intraclass correlation coefficient (ICC = ρ)	0 (no interdependence) to 1 (complete interdependence)	Less than 0.01 - 0.02 in sample sizes greater than 100.	The interdependence among the individuals within a cluster. The lower the ICC the better.
Design effect (DE)			A numeric correction factor introduced by the investigator to adjust for the presence of clusters in the calculation of the effective sample size. $DE = 1 + \rho (m - 1)$

Table 2
Power calculation while holding number of clients constant (m is constant)**

Number of Salons (k)	Number of Clients per salon (m)	Total N in study (mk)	ICC = $\rho = .025$	Power T-test*
			DE	ESS
Power calculation while holding number of clusters constant (k is constant)				
21	70	1470	2.725	.540
21	50	1050	2.225	.472
21	30	630	1.725	.365
21	1	21	1	.21
Power calculation while holding number of clusters constant (m is constant)				
25	50	1250	2.225	.562
21	50	1050	2.225	.582
15	50	630	2.225	.337
1	50	50	2.225	.112
Power calculation while holding total N constant (mk is constant)				
1	1050	1050	27.225	.39
10	105	1050	3.60	.292
21	50	1050	2.225	.472
30	35	1050	1.85	.568

ρ (rho) is a hypothetical intracluster coefficient (ICC)

DE (design effect) = $1 + \rho (m - 1)$

ESS (effective sample size)= mk / DE

* power of a 2 sample t-test, with $p < .05$ and a small effect size

** data presentation in table modeled from Killip et al., 2004 article

Table 3

Holding total N constant with varying intraclass correlation coefficient (ρ).

Number of Salons (k)	Number of Clients per salon (m)	Total Number (mk)	DE	ESS	ρ	Power T-test [*]
21	50	1050	3.45	304	.05	.411
21	50	1050	2.225	472	.025	.582
21	50	1050	1.49	705	.01	.755

Total N in study (mk)

DE (design effect) = $1 + \rho (m - 1)$

ESS (effective sample size)= mk / DE

ρ (rho) is a hypothetical intraclass correlation coefficient (ICC)

* power of a 2 sample t-test, with $p < .05$ and a small effect size