Advanced Statistics: Statistical Methods for Analyzing Cluster and Cluster-randomized Data

ROBERT L. WEARS, MD, MS

Abstract. Sometimes interventions in randomized clinical trials are not allocated to individual patients, but rather to patients in groups. This is called *cluster allocation*, or *cluster randomization*, and is particularly common in health services research. Similarly, in some types of observational studies, patients (or observations) are found in naturally occurring groups, such as neighborhoods. In either situation, observations within a cluster tend to be more alike than observations selected entirely at random. This violates the assumption of independence that is at the heart of common methods of statistical estimation and hypothesis testing. Failure to account for the dependence between individual observations and the cluster to which they belong can have profound implications on the design and analysis of such studies. Their p-values will be too small, confidence intervals too narrow, and sample size estimates too small, sometimes to a dramatic degree. This problem is similar to that caused by the more familiar "unit of analysis error" seen when observations are repeated on the same subjects, but are treated as independent. The purpose of this paper is to provide an introduction to the problem of clustered data in clinical research. It provides guidance and examples of methods for analyzing clustered data and calculating sample sizes when planning studies. The article concludes with some general comments on statistical software for cluster data and principles for planning, analyzing, and presenting such studies. **Key words:** statistics; cluster data; research; clinical trials. ACADEMIC EMERGENCY MEDICINE 2002; 9:330–341

SOMETIMES interventions in randomized clinical trials are not allocated to individual patients, but rather to patients in groups. This is called *cluster allocation*, or *cluster randomization*, and is particularly common in health services research, but can be found in many other settings as well. There are at least three reasons why designers of studies might wish to allocate subjects to interventions in groups.

First, the intervention may naturally be applicable to the cluster. For example, consider a controlled trial of an educational intervention directed at physicians to reduce inappropriate antibiotic prescribing. It would be impossible to assign patients to intervention and control groups, since a physician has either received the educational intervention or has not, and its effect if received cannot be "turned off." Thus, the only alternative is to allocate physicians to the intervention, even though the effect is measured on prescriptions given to patients.

Second, even if individual allocation is possible, there may be significant crossover contamination if individual allocation is used. If the physicians in the preceding example worked together in group practices, it is likely that the physicians in the control group, who did not receive the intervention, might still be affected by it via conversations with their colleagues in the intervention group. Thus, allocation might be better done at a higher level, the level of the practice. As another example, consider a controlled trial of computer-based order entry to reduce medication errors (or cost, inappropriate prescribing, etc.). In principle, it might be possible (although impractical) to assign patients to be managed by computer-based or traditional written orders, but it is highly likely that the physicians' ordering decisions would be affected by their experience with the computer-based system. Educational interventions offered to patients waiting to be seen in emergency departments (EDs) are a third example. Although patients could be randomized, they would likely discuss their experiences with other patients in the waiting room. Thus, allocation at the level of the ED might be desirable in these cases. (However, cluster allocation is not the only method of handling contamination, and may not always be the best approach to take.[1])

Third, sometimes cluster allocation is much less

From the Department of Emergency Medicine, University of Florida Health Center (RLW), Jacksonville, FL.
Received May 29, 2001; revision received October 10, 2001; accepted December 10, 2001.
Series editor: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor–UCLA Medical Center, Torrance, CA.
Presented as a didactic session at the SAEM annual meeting, San Francisco, CA, May 2000.
Address for correspondence and reprints: Robert L. Wears, MD, MS, Department of Emergency Medicine, University of Florida Health Center, 655 West 8th Street, Jacksonville, FL 32209. Fax: 904-244-4508; e-mail: wears@ufl.edu

costly or more practical than individual allocation. In the computer-based ordering example, allocation by individual would require the computer-based system to be installed at every site, doubling the cost, but be used on only half the patients at each site, a highly inefficient design. Many public health interventions are relatively less costly when implemented at an organizational level (say, the county) than at an individual level.

An analogy to cluster randomization/allocation occurs in observational studies. Here it can sometimes be more efficient to gather data from organizational units, such as blocks, census tracts, or counties, than from individuals. In addition, sometimes patients can choose which cluster they belong to, e.g., they can choose to use one health plan and not another, or one ED and not another. Two patients within a cluster might therefore be more similar to one another than two patients randomly chosen at large, so it would be reasonable to take this into account in both sampling and analysis. In this paper, for simplicity's sake, we will speak of cluster randomization, but the same principles will apply to cluster allocation and to cluster sampling, unless specifically noted.

A final situation that is closely related to cluster sampling is that of profiling data. Here data are collected, for example, on physician or hospital performance, frequently for the purpose of ranking or identifying outliers so they might be persuaded or cajoled to return to the fold. If outcomes within a hospital or physician (i.e., a cluster) tend to be more consistent than outcomes among hospitals or physicians, the standard error will be underestimated if traditional statistical methods are used, and outliers will consequently appear to perform better or worse than they really are.[2]

Although in this paper we will consider only a single level of clustering, clustering can occur at multiple levels. For example, patients might be clustered under the ambulance crews who treat them, while ambulance crews might be clustered under the municipal departments to which they belong, as shown in Figure 1. This is an example of hierarchical clustering, where each lower level unit belongs to one and only one higher level unit. A more complex situation of non-hierarchical clustering is shown in Figure 2. Here patients are clustered under ambulance crews as in Figure 1, but patients also belong to health plans, which are not unambiguously associated with ambulance crews. Extensions of the methods described here can handle multiple levels and types of clustering.

What Is the Problem with Cluster Data?

When subjects are randomized, allocated, or sampled by cluster, it creates several problems for statistical analysis. Observations within a cluster tend to be more alike than observations selected entirely at random. There are at least three reasons why this is so.[3] First, in some situations patients may choose which cluster they belong to. For example, patients choose one ED and not another, one physician and not another; residents choose one training program and not another. It seems likely that people choosing the same cluster may also be similar in other ways. Second, cluster-level variables may affect all members simultaneously. For example, patient outcomes from a given procedure may differ among different physicians. Third, individuals within clusters may interact and influence each other. Patients in an ED waiting room may share information and thus influence each other's behavior.

If observations in a cluster are correlated, then they are not statistically independent, violating one of the fundamental assumptions of estimation and hypothesis testing. If we cannot think of the

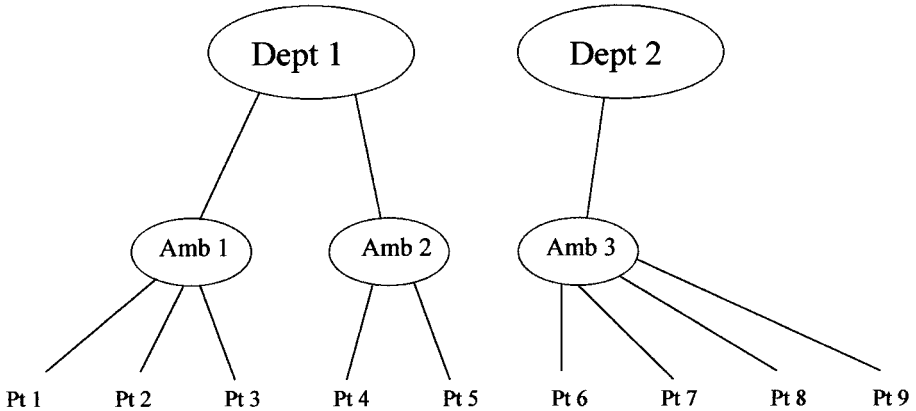

Figure 1. Example of hierarchical clustering. Patients (Pt) are clustered under ambulance crews (Amb), who in turn are clustered under public safety departments (Dept). Knowing which ambulance crew treated a patient unambiguously determines which department was involved.

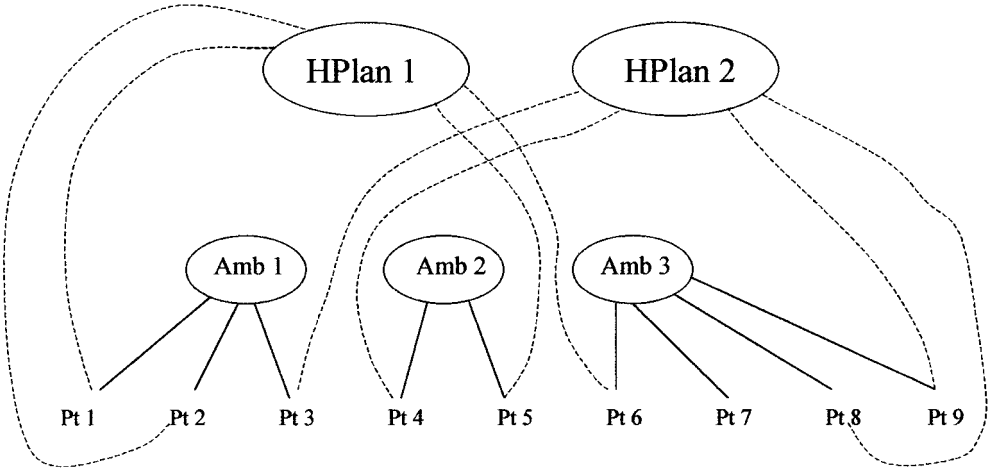

Figure 2. Example of nonhierarchical clustering. Patients are clustered under ambulance crews as in Figure 1, but they are also clustered under health plans (HPlan), shown by the dotted lines. Knowing which ambulance crew treated a patient does not determine his or her health plan, and vice versa. (Patient 7 is uninsured and belongs to no health plan.)

observations as independent, we must modify the analysis to take the cluster design into account.[4] When cluster designs are used, there are two sources of variance in the observations. The first is the variability of patients within a cluster, and the second is the variability between clusters. These two sources combine to produce an increase in the variance (relative to the situation of a single cluster), and both must be taken into account in the analysis.

The effect of the increased variance due to a cluster design is to increase the size of the standard errors and thus to widen confidence intervals and increase p-values, compared with a study of the same size using individual randomization. In effect, the sample size is reduced and power is lost. Conversely, failing to account for clustering in the analysis will result in confidence intervals that are falsely narrow and p-values that are falsely low; in other words, the risk of a false-positive error is increased. The magnitude of the increase may be substantial, and can approach 50%. In the words of Jerome Cornfield, "Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception."[5] Interestingly, Cornfield wrote these words in 1978, but issues of cluster randomization have only been noted in general biomedical journals in the last few years.

It may be easier to understand the problems associated with cluster designs by drawing an analogy to another, perhaps more familiar problem, the "unit of analysis error."[6] In its classic form, the unit of analysis error occurs when multiple observations are taken on the same subject, but are treated as if they were independent. For example, a patient may have multiple measurements, say of joint mobility, or may have the same measurement (e.g., blood pressure, or BP) repeated over time. It would be wrong in these situations to regard the separate data points as independent, since measurements taken on the same patient are more likely to be related than measurements taken on different patients. In addition, the sample size is artificially inflated if the data are analyzed at the observation level. For example, if five measurements of BP are taken on 10 patients, treating the BPs as independent observations gives a sample size of 50, but in reality, it is only 10. Here the patients, not the separate BP measurements, are the appropriate choice for the unit of analysis; analogously, one could say that each patient in this example is a cluster.

Just as BP measurements on a single patient are more likely to be similar than if taken on different patients, measurements taken on individuals in a nonrandomly formed cluster are more likely to be similar than those taken on individuals at random. For example, residents within a given residency program are likely to be more similar in most aspects than residents selected at random across the country. Acting as if these observations were independent is the same kind of mistake as the unit of analysis error, for example, treating multiple BP measurements on the same patient as if they were independent.

We can gain some insight into the nature and magnitude of the problem using a little algebra.[7] Suppose we measure some variable y on n individuals who have been divided into k equal groups, or clusters. (The measurements can be continuous or discrete, and need not be normally distributed.) The clusters need not be equal, but for simplicity we will assume that here, and so each cluster contains $m = n/k$ observations. If the cluster assignments were random, then the errors in measure-

ment on members of a cluster are independent, and can be expressed by the within-cluster variance, σ_y^2. The variance of the kth cluster's mean would then be:

$$\sigma_{\bar{y}_k}^2 = \frac{\sigma_y^2}{m} \quad (1)$$

If the variance were constant across all the clusters, then the variance of the "grand mean" would be:

$$\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{mk} = \frac{\sigma_y^2}{n} \quad (2)$$

This variance is the one that would be used in computing hypothesis tests or confidence intervals when describing this group or comparing it with some other group or groups, and is the variance one would obtain from an individual-level analysis.

Now suppose that the individuals were not randomly assigned to the k clusters, but had somehow aggregated into k identifiable clusters through some unknown nonrandom process. Members of an identifiable cluster will have something in common, due to similarities in selection, exposure to outside influences, interaction with one another, etc. Because the cluster members are similar, the variance within a cluster will be smaller than it would have been for randomly assigned members. The degree to which the within-cluster variance σ_e^2 is smaller than would otherwise be expected can be expressed in terms of the intracluster correlation coefficient (ICC), ρ, as follows:

$$\sigma_e^2 = \sigma_y^2(1 - \rho) \quad (3)$$

Equation 3 shows that if the intracluster correlation is high, then the within-cluster variance will be small. If the intracluster correlation is zero (i.e, the observations are competely independent), then the within-cluster variance is equal to the ordinary variance. The ICC can be interpreted as the proportion of the total variance in the data that is due to the clusters, or:

$$\rho = \frac{\sigma_k^2}{\sigma_y^2} = \frac{\sigma_k^2}{\sigma_k^2 + \sigma_e^2} \quad (4)$$

Although in theory the ICC could be negative, in practice with naturally identifiable clusters, that almost never occurs. (However, it can create some problems with certain estimation methods.) Now that we can separate the cluster and within-cluster contributions to the total variance, we can write the variance of the kth cluster's mean as the sum of the two-variance components:

$$\sigma_{\bar{y}_k}^2 = \frac{\sigma_e^2}{m} + \sigma_k^2 \quad (5)$$

With some algebraic manipulation, equation 5 can be rewritten in a more useful (believe it or not) form,[8] as:

$$\sigma_{\bar{y}_k}^2 = \frac{\sigma_y^2}{m}(1 + (m - 1)\rho) \quad (6)$$

From this it follows that the variance of the grand mean will be:

$$\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{mk}[1 + (m - 1)\rho] \quad (7)$$

Equation 7 shows clearly that the variance of the overall mean in cluster data is equal to the ordinary variance multiplied by some "variance inflation factor" (the quantity $1 + (m - 1)\rho$). This factor is called the "design effect" by some authors,[9–11] but other sources use a slightly different definition of the design effect.[12] We will use the term variance inflation factor (VIF) throughout this paper so that we can distinguish the two definitions when appropriate. (In practice, the two definitions provide estimates that are generally comparable in magnitude.) The VIF is an extremely useful entity. It is the factor by which the total sample size must be increased if a cluster design is to have the same statistical power as an individual design.[11] It can also be used in statistical inference to correct confidence intervals and p-values.

If the cluster sizes are not equal, which will commonly be the case with real data, the average cluster size should not be used; instead, an adjusted mean cluster size, m', is used, calculated from[3]:

$$m' = \frac{1}{k - 1}\left(n - \frac{\sum_k m_k^2}{n}\right) \quad (8)$$

If the ICC is greater than zero, then several important implications follow from equation 7. First, the variance in a cluster design will always be greater than an individual design. This means that, all else being equal, the confidence intervals will be wider and the p-values larger.

Second, if the cluster size m is large, even very small values of the ICC can have important effects. This is quite different from our usual approach to correlation coefficients, where small correlations can be generally ignored. If the average cluster size is 50, an ICC of 0.02 would give a VIF of 2, implying that the cluster sample needs to be twice as large as that calculated for individual sample

TABLE 1. Results of Simple t-tests That Do Not Account for the Cluster Design

	Intervention	Control	95% CI on Difference	p-value
Score1	74.8	74.9	−0.58 to +0.45	0.80
Score2	78.5	74.3	1.15 to 7.17	0.008
Delta	3.67	−0.55	1.09 to 7.35	0.009

design in order to obtain the same statistical power.

Third, any statistical test that ignores clustering will have a type I error rate greater than its nominal level. That is, if the test cutoff is set at 0.05, the probability of a type I error will actually be greater than 0.05 by an amount determined by the size of the VIF.

Does It Matter? So far, this may have seemed highly theoretical, a "distinction without a difference." Does cluster allocation or cluster sampling really matter in the real world, or is it simply one more thing for statisticians to use to intimidate others? The answer, of course, is that it can matter a lot.

The problem of inappropriate analysis of cluster data appears to be fairly prevalent. Divine et al. reviewed 54 studies on physicians' behavior from a broad selection of journals and found 70% used the wrong unit of analysis.[13] They also reanalyzed data from another study, this time using the physician as the (correct) unit of analysis instead of the patient. In the original, incorrect, by patient analysis, eight of nine measures were statistically significant, but in Divine et al.'s reanalysis by physician, only four were significant. A similar study of 21 public health papers showed that only 57% accounted for clustering in their analysis. Of the nine trials that failed to account for clustering, it appeared that the conclusions were spuriously positive in seven. Less than 20% of the studies took clustering into account in calculating sample size or power.[9] Of the 17 trials that failed to do so, three found no effect, possibly as a result of lack of statistical power. Several other reviews have found results similar in magnitude.[14,15]

ANALYZING CLUSTER DATA

Since we have seen that applying traditional statistical methods to cluster randomized observations as if they were individually randomized is wrong, what alternate approaches are available? There is a rich literature on the analysis of complex designs, which can be almost overwhelming. Much of the published material is aimed at designs much more complicated than those discussed here, and include problems such as multiple levels of clustering, which can be either hierarchical or nonhierarchical, probability sampling, stratification, matching, adjustment for covariates, and time trends, to mention only a few. In addition to complexity, much of this work was originally done for social science problems (consider students within classrooms within schools), or for public health/epidemiologic problems, so emergency physicians would be relatively unlikely to be exposed to this literature. Finally, the field suffers from lack of a common nomenclature, so that "groups" and "clusters" sometimes have different meanings in different papers. Variables in published formulas frequently have different definitions, e.g., n can sometimes refer to the total number of patients, the total number of clusters, or the number of patients in a cluster.

The goal of this paper is to provide a gentle introduction to the analysis of cluster data, using three basic approaches: traditional analysis by allocation unit, individual analysis using adjusted variance estimates, and multilevel modeling methods. These are not the only methods useful for cluster data and are not always the best choices, but will give a flavor for the correct approach and should assist the reader in tackling more complex methods. (More detailed information on a variety of methods, including Bayesian approaches,[16,17] can be found in *Statistics in Medicine* 2001; 20(3), which dedicates an entire issue to these problems.) In particular, a conscious attempt has been made to use a single nomenclature and notational system throughout the paper. Finally, it should be noted that cluster designs present many more problems than simply that of choosing the correct analytic method. Cluster randomization is more prone to bias than individual randomization, and careful attention must be paid to identifying, eliminating, or at least measuring a number of sources of bias in the design and execution of cluster-randomized trials.[7] In addition, cluster randomization can raise difficult ethical and regulatory questions about consent whose answers are not yet well worked out.[18–20]

Example Analyses. The data set given in Appendix A will be used to outline some of the basic analytic methods. (A downloadable version of the data set and STATA 7.0 code to reproduce the analyses in this section are available on the *Academic Emergency Medicine* web site, aemj.org). This hypothetical data set represents an experiment on educational methods carried out at six different residency programs as a cluster-randomized trial. The six programs (represented by the *center* variable) were randomized to an intervention (*group* = 0) and a control group (*group* = 1). Eight residents in each program were tested before and after the

intervention (*score1* and *score2*). The difference between pre- and postscores (*delta*) was the primary outcome variable. A value for delta greater than zero represents an improvement, and the null hypothesis is that the mean deltas for each group are equal. Because the investigators thought experience might be an important covariate, they also recorded each resident's number of years of experience in the *yrs* variable. The *pass* variables represent before and after "passing" rates, defined as a score greater than 75. We will examine several possible ways of analyzing these data. (This analysis uses STATA, version 7.0. Other software should give similar, but perhaps not exactly the same results.)

Analysis by Individual (*wrong*!). Since the intervention was allocated by cluster, analyzing these data as if the residents were individually randomized would be wrong. We will compare the conclusions reached by this erroneous analysis with better, alternate approaches.

Simple t-tests on score1, score2, and delta that do not account for the cluster design show the group means to be roughly comparable at the start, and that the intervention group shows a significant improvement while the control group stays roughly the same (Table 1). However, this conclusion is suspect because it failed to account for the cluster design. Of course, hypothesis testing on score1 is not a good way of establishing equivalence in a randomized trial, and generally should not be done.[21] When the number of clusters in a cluster-randomized trial is small (less than 10 to 20), randomization may not ensure balance on unmeasured covariates. Even then, hypothesis testing on baseline characteristics is not a good strategy; stratification, matching, and statistical modeling are better choices.

Analysis of the proportions passing before or after is similar. The pass rates were not significantly different before the intervention (42% and 38%, p = 0.77), but favored the intervention group afterwards (42% and 75%, p = 0.02).

Analysis by Allocation Unit. The simplest alternative to individual analysis is to simply reduce all the individual observations within a cluster to a single summary measure, such as the cluster mean or proportion, and then use standard statistical methods to analyze these summary measures as if they were the primary observations. This entails a drastic reduction in sample size, and consequently statistical power. The center means for delta by group are given in Table 2.

TABLE 2. The *Center* Means for *Delta* by Group

	Intervention	Control
	1.02	1.37
	6.06	−2.94
	3.93	−0.09
Grand mean	3.67	−0.55

A t-test on the difference in means gives t = 2.19. This would be significant if the original number of degrees of freedom (46) could be used, but the sample size here is only 6, so for a two-group t-test, the number of degrees of freedom is 4, for p = 0.09. The difference in means is 4.22, with a 95% confidence interval on the difference of −1.14 to 9.58, about 35% wider than the interval calculated by the erroneous method above. Notice that the difference in using a better analytic method does not affect the estimates of the means, but profoundly affects the variance, which in turn affects the widths of the confidence intervals and p-values. The proportion measures (*pass1* and *pass2*) are similarly affected. Analysis by individual has led to a false-positive conclusion.

This method has the advantage of not requiring specialized software tools, but at the expense of a profound reduction in sample size and degrees of freedom. However, when the number of clusters is small (less than ten), estimation of the ICC and modeling methods may be unreliable, so this method may be the only reasonable choice. It is known to be robust (at least for the t-test) with as few as three clusters.[3] It also suffers from vulnerability to the *ecological fallacy*, discussed below, although it is possible to overcome some of the inability to adjust for individual-level covariates by standardizing the cluster level summary statistics (means, proportions, odds ratios) for important covariates.

This method gives equal weight to all clusters. In the example, the cluster sizes are equal, but in real data they typically are not, so some method of weighting the cluster means for analysis is frequently desirable. While weighting by cluster size is intuitively appealing, it may lead to reduced power and other problems in the analysis. Inverse variance weighting has been proposed to reduce these problems and increase statistical power (or lower sample size requirements); it also provides a more precise estimate for the grand mean.[22] This is roughly analogous to performing inverse variance-weighted random effects meta-analysis on the data, pooling the summary results across clusters rather than across studies.[23,24]

Analysis Using the VIF. Another approach[7] also uses standard statistical methods applied to individual data, but uses the VIF to correct the variance used in calculation of the test statistic. For χ^2 and F tests, the denominator of the test statistic is a variance, so dividing the calculated test statistic by the VIF results in a revised test statistic that

will be approximately correct if the VIF can be estimated reliably. For t-tests, the denominator of the t statistic is the square root of the variance, so the value for t is divided by the square root of the VIF. These revised statistics must still be referred to the appropriate number of degrees of freedom based on the number of clusters, not on the total sample size.

For our sample data, the ICC for delta is estimated to be 0.1877. Not all statistical packages easily produce an ICC (or intraclass correlation coefficient). However, it can be calculated from the results of one-way analysis of variance (ANOVA) as shown below; Fleiss gives a derivation[25]:

$$ICC = \frac{BMS - WMS}{BMS + (m' - 1)WMS} \tag{9}$$

where *BMS* is the mean square between clusters, *WMS* is the mean square within clusters, and m' is the mean cluster size (or if the clusters are not equal, the adjusted mean cluster size as calculated in equation 8). In this case the VIF for an ICC of 0.1877 and a cluster size of eight is 2.31. The t statistic for a difference in two group means is given by:

$$t = \frac{x_2 - x_1}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{10}$$

where s_p^2 is the pooled estimate of variance. The t statistic can be corrected for the cluster design by multiplying s_p^2 by the VIF; this is equivalent to dividing t by the square root of the VIF. For the t test on delta, calculated for individual randomization, t is 2.713, and the corrected t would then be 1.784. This value is referred to 4 degrees of freedom, to yield p = 0.15. Confidence intervals are calculated in a similar manner using the adjusted variance, and for these data yield a 95% confidence interval of −2.34 to 10.79, somewhat more conservative (in this instance) than that calculated above, but qualitatively in agreement. The value of this method is that it allows "after-the-fact" correction for published p-values and confidence intervals if reasonable estimates for the intracluster correlation are available.

If the group sizes, variances, numbers of clusters, or cluster sizes are appreciably different, the formulas given above should be modified to calculate the VIF separately for each group, and to adjust each group's variance separately when estimating the combined variance.[3]

Multilevel Methods. Both "cluster-aware" methods outlined above suffer from vulnerability to the ecological fallacy, the error that can result from applying attributes of groups to their members. (An example would be concluding that Protestants are more likely to commit suicide than Catholics from the observation that suicide rates are higher in Protestant countries. It could be the Catholics in those countries, alone and surrounded by gloomy Protestants, who are committing suicide!) Thus, it would be desirable to be able to adjust the analysis for individual-level covariates, even when conducting the analysis at the level of the cluster. There are a variety of methods, generally termed multilevel, hierarchical, or random-effects methods, that can be used for this. They are based on a wide variety of statistical models, such as the generalized linear mixed model, generalized estimating equations, and hierarchical Bayesian models. These models can be viewed as extension to the general linear regression model.

The general linear regression model for a single covariate can be summarized as follows:

$$y_{ij} = \alpha + \beta x_{ij} + e_{ij} \tag{11}$$

where y_{ij} is the outcome for the ith individual in the jth cluster, α is the intercept or constant, β is the slope or regression coefficient that describes the relationship between x and y, and e_{ij} is a random error term with zero mean and constant variance. The hierarchical or multilevel model can be similarly summarized as:

$$y_{ij} = \alpha + \beta x_{ij} + \mu_j + e_{ij} \tag{12}$$

where μ_j is an additional term modeling the random effect of the jth cluster, and measures the amount by which the jth cluster differs from the overall mean. It should be noted that the coefficients for hierarchical models will not be equivalent to those obtained by running a separate regression on each cluster, because they use information from all the observations in their estimate. In effect, cluster-specific coefficients are shrunk toward the mean for the overall model. A major drawback to these methods are that they require a relatively large (~25) number of relatively large (~25 member) clusters to achieve their asymptotic performance. This effectively limits them to observational data for the most part.

We will use the generalized mixed linear random effects model provided in STATA to illustrate this method on our example data. Using the *center* variable to identify the cluster (or primary sampling unit), STATA's estimate of the design effect is 2.38, not substantially different from our previous estimate of the VIF. The design-adjusted estimates of delta for the intervention and control groups are 3.67 (95% CI = 0.32 to 7.02) and 0.55 (95% CI =

−3.46 to 2.36). The estimate for the difference in means is 4.22, as before, with a confidence interval from −0.22 to 8.66, roughly 50% wider than the individually-based estimate. The t statistic is 2.445, which with 5 degrees of freedom corresponds to a p-value of 0.058. (Note that the use of a design-based test has resulted in a gain of 1 degree of freedom. With large numbers of clusters, this might not be significant, but with small numbers of clusters as in this example, the difference could be important.) Similarly, tests on the proportion passing after the intervention showed a difference favoring the intervention group of 33%, as before, but now with a 95% confidence interval of −5% to +72%, and p = 0.08.

The advantage of using a hierarchical model is that it can be extended to handle stratification as a means of controlling for confounding and reducing bias, or regression adjustment for important covariates (confounders, effect modifiers, or nuisance variables). In this example, we might want to adjust for the effects of different levels of experience on the results. One way to do that would be with a regression model using delta as the dependent variable, and group and years as independent variables. We do not use center in the regression equation because it has been identified as the primary sampling unit and is accounted for there (Table 3).

The coefficient for group is negative; since our coding set the intervention equal to 0 and the control equal to 1, this means that the intervention is associated with an improvement in delta of about 3.25. Also, including experience in the analysis now looks like a good idea, since it appears to have been confounded with group. One additional year of experience has an effect almost as great as that of the intervention.

At this point, one might wonder why simply including center as a nested categorical variable in a linear model is not a reasonable strategy for analysis. The answer is that that could be, but only if the analyst can correctly specify the fixed and random factors and calculate the variances accordingly. Most statistical packages default to fixed effects, or if they allow some random effects, do not clearly specify how to calculate error terms. For example, SAS PROC GLM presumes a fixed-effects model; although it allows the user to designate some effects as random, it has a type I error rate close to the nominal rate only if the data are balanced and there is no appreciable confounding.[7] Since the cluster(s) will always be random effects, it is best not to assume that the software will know how to construct error terms and F tests correctly. In addition, the number of clusters required for reliable performance is large, and each cluster consumes an additional degree of freedom.

TABLE 3. Regression Model Using *Delta* as the Dependent Variable and *Group* and *Yrs* as Independent Variables

Variable	Coefficient	Standard Error	p-value
Constant	−2.72	1.50	0.13
Group	−3.25	1.20	0.04
Yrs	2.60	0.69	0.01

Summary of Analytic Methods. Any time there are multiple analytic options, there will be differences of opinion among statisticians about which choice is best, although most will admit that previous familiarity with a method plays a large role in their opinion. This uncertainty is compounded when there is not a "natural" choice for the analytic unit. For example, consider a school-based intervention aimed at smoking prevention. The observations would be taken on individuals, who are nested within classes, which are nested within teachers, who are nested within schools, which are nested within school districts. A plausible case can be made for analysis at several of these different levels. However, there are some general points of consensus.

Most analysts agree that the allocation unit should be the unit of analysis.[7] There is broad agreement that, with the possible exception of Bayesian methods, multilevel methods are more reliable when the number of clusters is relatively large, e.g., greater than 25. Methods based on adjustment by the VIF require stability in the estimation of the ICC, which seems to require 10−20 clusters. Studies based on fewer than 10 clusters would be most conservatively handled by analysis at the cluster level, possibly supplemented by some cluster-specific adjustment of means. There are many special approaches that may be useful in context, and researchers would do well to consult with a statistician experienced with these methods.

Sample Size Calculations for Cluster Designs. Since the variance is affected by cluster designs, the sample size required for a certain power and effect size is also affected. Because the sample size is directly proportional to the variance, we can simply use the standard methods and multiply the result by the appropriate VIF, calculated on the basis of a presumed value for the ICC and estimated average cluster size. For our hypothetical example, suppose we wished to have 90% power to detect a difference in delta of 5 points at the 0.05 level. We believe the standard deviation of delta is about 5, the average cluster size will be about 8, and that the ICC will be about 0.2. The VIF will therefore be 2.4. Standard sample size calculations show that 23 patients per group, or 46 overall, will

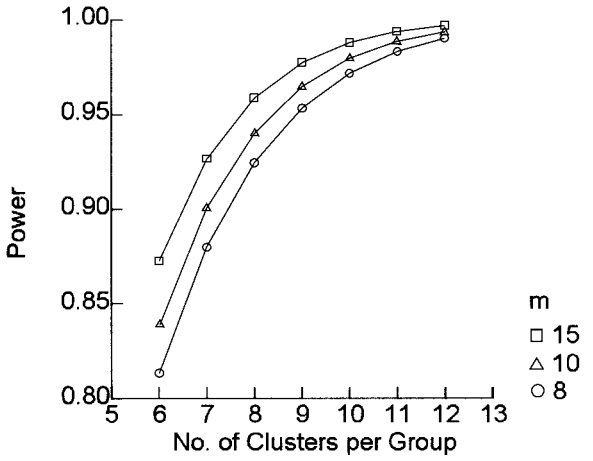

Figure 3. Estimates of statistical power for a two-group t-test as a function of the number of clusters per group, and the cluster size, m. Note that since there are two groups, the total number of clusters, k, is twice the number shown on the x-axis.

be necessary, if individual randomization is used. Multiplying that by the VIF of 2.4 gives a total sample size of 110 patients in about 14 clusters. A further adjustment might be necessary to account for the loss in degrees of freedom. For an allocation unit analysis, or a VIF-adjusted analysis, there will be only 12 degrees of freedom for the t-test. The critical value of $t_{12,0.05}$ is 2.18, so increasing the sample size by the ratio of 2.18 to the critical value for the normal distribution (1.96) suggests increasing by a factor of 1.11 to a total size of 122 patients in roughly 15 clusters.

Although increasing the number of clusters is a more efficient method of gaining statistical power than is increasing total sample size, it may sometimes be easier to add observations to existing clusters rather than adding more clusters. This requires calculating a range of potential VIFs for different cluster sizes. One commercial sample size product (PASS 2000, from NCSS, Kaysville, UT) supports sample size and power estimation for cluster randomized trials. Figure 3 shows a plot of the output of PASS 2000 for the situation above. It is easy to see that a study with eight clusters of size 8 per group (total of 16 clusters and 122 individual) has roughly the same power as a study with seven clusters of size 15 per group (total of 14 clusters and 210).

Clusters vs Strata. A common problem for many when first considering cluster data is the distinction between clusters, which require special analytic methods, and strata, for which well-known methods exist. The distinction is analogous to the discussion of fixed vs random effects, above. Strata are fixed effects, in that the investigator is not interested in generalizing beyond the values of the stratification variable contained in the data. Another way of saying this is that if the data contain all the values of interest for a variable, then it is a fixed effect and can be handled as a stratification variable. For example, gender would be a stratifying variable and not a cluster variable, even though there may be similarities in response within gender, because the data contain all the values of gender to which the inferential results will be applied, i.e., male and female. Neighborhood, or zip code, on the other hand, would be a cluster variable, not a stratifying variable, because the data do not contain all the values for zip code to which the inferences are intended—the results are of interest only if they can be applied to all neighborhoods, not just the specific ones that happened to be included in the sample.

Software. A number of commercial software packages support analysis of cluster data in one way or another. A variety of "specialist" packages in various states of maintenance and development are also available.[3,7] A noncomprehensive list of commercially available packages,[7] with some gratuitous comments, includes the following:

- SAS (SAS Institute, Cary, NC; http://www.sas.com). PROC MIXED is a general mixed-model regression program that supports multiple random effects. PROC GLM should not be used; it is a fixed-effects model and has been shown to have a highly inflated type I error rate.[7] SAS can also be used to implement a generalized estimating equations approach.
- Stata. (Stata Corporation, College Station, TX; http://www.stata.com). Stata provides random effects modeling in its *svy* procedures. Stata uses a "robust" estimator that has few assumptions, but may perform poorly when the number of clusters is small (e.g., less than 10 clusters per group). Stata also supports generalized estimating equations.
- EGRET (Cytel Corporation, Cambridge, MA; http://www.cytel.com). Provides random effects modeling.
- SUDAAN (Research Triangle Institute, Research Triangle Park, NC; http://rti.org/sudaan). SUDAAN was designed specifically for analysis of complex surveys, and can handle many complex designs including multilevel clustering. It uses several different "robust" estimators.
- BMDP (Statistical Solutions, Cork, Ireland; http://www.statsol.ie/index.html). Provides a generalized mixed-model regression facility.
- BUGS/WinBUGS (The BUGS Project, Cambridge, UK; http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml). Supports hierarchical Bayesian modeling using computer-intensive estimation methods. Should be considered experimental, but

can provide estimates unavailable in other packages.

Final Recommendations

Although they are more problematic than individual randomization or sampling, it should be clear that cluster randomization and cluster sampling are sometimes the only option (or the only practical option) available for addressing some types of questions. When the mechanism by which individuals are assigned to clusters is not random (for example, they can select which cluster they belong to, they may interact and thereby influence one another, or they may be subject to the shared influence of cluster level effects), a design based on the individual is inappropriate because of the potential for contamination.[26–28] However, when a cluster design is used, it is important that the analysis address clustering appropriately.[29] The following recommendations, adapted from Ukoumunne et al.,[26] summarize for authors and trialists the issues that must be addressed in designing, analyzing, and reporting these studies.

1. Recognize the cluster as the unit of intervention or sampling. The analysis should be carried out on the same unit to avoid spurious results.

2. Justify the use of the cluster as the allocation or sampling unit. The reasons for choosing a cluster design should be explicitly stated. Theoretical and practical reasons are both acceptable in appropriate circumstances. Examples of good reasons might include:

a. The intervention is naturally applied at the cluster level (e.g., when educational interventions are directed at physicians or physician practices).

b. Contamination is likely with individual allocation (e.g., due to interaction among individuals).

c. Cost may be lower in cluster allocation.

3. Include enough clusters. Power in a cluster design depends much more on the number of clusters than on the total sample size. Studies with only two clusters are in principle unable to distinguish between cluster effects and intervention effects, since they are completely confounded. As a general rule, there should be at least four clusters in order to have even minimal statistical power, and often many more will be needed. Eight clusters of 25 patients (total 400) will be more powerful than four clusters of 50 patients (total 400). Some analytic methods may be unreliable when the number of clusters is low, so the analysis plan should be adapted to the number of clusters.

4. Randomize clusters whenever possible to avoid bias. If randomization is not possible, then before–after measurements in intervention and control groups should be used to at least partially adjust for group differences.

5. Allow for clustering when estimating the required sample size. Standard sample size formulas will underestimate the number of observations required because they allow only for variation within clusters but not between clusters. The sample size derived from standard formulas should be multiplied by the design effect or variance inflation factor to give the cluster level analysis the same power to detect a given effect magnitude as a study with individual allocation and analysis. Increasing the number of clusters will provide greater increases in power than will increasing the number of patients within clusters, and will also enhance generalizability.

6. Allow for clustering in the analysis. Performing an individual-level analysis on cluster data will result in confidence intervals that are too narrow and p-values that are too small. The analysis can be conducted at the cluster level by applying standard methods to the cluster means or proportions, or at the individual level using special methods such as multilevel models or generalized estimating equations that can take into account the similarity of observations within a cluster.

7. Allow for confounding at the individual and cluster levels. Analysis at the cluster level is vulnerable to the ecological fallacy (the attribution of group characteristics to individuals). If there are important confounding variables, stratification, matching, or regression models for clustered data are required. Multilevel models explicitly model the association of observations with clusters, while generalized estimating equations treat it as a nuisance variable. Both methods may require a fairly large number of clusters.

8. Include estimates of intracluster correlation and components of variance in published reports. Planning cluster-randomized trials is currently hampered by the paucity of information about the magnitude of the ICC in common circumstances. Publication of values of the intracluster correlation coefficient will help in planning future studies by providing at least an order of magnitude estimate of the ICC.

Finally, extensions to the CONSORT[30] statement to adapt it to cluster randomized trials have recently been suggested.[31] While these have not yet been formally adopted by biomedical editors, they should serve as useful guides for designers, readers, and reviewers of cluster randomized trials.

References

1. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ. 2001; 322:355–7.

2. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. Ann Intern Med. 1997; 127:764–8.

3. Ukoumunne O, Gulliford M, Chinn S, et al. Methods for evaluating area-wide and organisation-based interventions in

health and health care: a systematic review. Health Technol Assess. 1999; 3:1–110.
4. Bland JM, Kerry SM. Statistics notes. Trials randomised in clusters. BMJ. 1997; 315:600.
5. Cornfield J. Randomization by group: a formal analysis. Am J Epidemiol. 1978; 108:100–2.
6. Altman DG, Bland JM. Statistics notes: units of analysis. BMJ. 1997; 314:1874.
7. Murray DM. Design and Analysis of Group-Randomized Trials. Oxford, UK: Oxford University Press, 1998.
8. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. Am J Epidemiol. 1981; 114:906–14.
9. 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.
10. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. BMJ. 1998; 316:1455.
11. Kerry SM, Bland JM. Sample size in cluster randomisation. BMJ. 1998; 316:549.
12. StataCorp. Stata Statistical Software: Release 6.0. College Station, TX: Stata Corporation, 1999.
13. Divine GW, Brown JT, Frazier LM. Unit of analysis error in studies about physicians' patient care behavior. J Gen Intern Med. 1992; 7:623–9.
14. 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.
15. Ennett ST, Tobler NS, Rignwalt CL, Flewelling RL. How effective is drug abuse resistance education? A meta-analysis of Project DARE outcome evaluations. Am J Public Health. 1994; 84:1394–401.
16. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. Stat Med. 2001; 20:453–72.
17. Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. Stat Med. 2001; 20:435–52.
18. Hutton JL. Are distinctive ethical principles required for cluster randomized controlled trials? Stat Med. 2001; 20:473–88.
19. Brett A, Grodin M. Ethical aspects of human experimentation in health services research. JAMA. 1991; 265:1854–7.
20. Edwards SJL, Braunholtz DA, Lilford RJ, Stevens AJ. Ethical issues in the design and conduct of cluster randomised controlled trials. BMJ. 1999; 318:1407–9.
21. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. Lancet. 1990; 335:149–53.
22. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med. 2001; 20:377–90.
23. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986; 7:177–88.
24. Thompson SG, Pyke SD, Hardy RJ. Design and analysis of paired cluster randomised trials: an application of meta-analysis techniques. Stat Med. 1997; 16:2063–79.
25. Fleiss JL. Reliability of measurement. In: The Design and Analysis of Clinical Experiments. New York: John Wiley & Sons, 1986, pp 1–32.
26. Ukoumunne OC, Gulliford MC, Chinn S, et al. Methods in health service research. Evaluation of health interventions at area and organisation level. BMJ. 1999; 319:376–9.
27. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlation for the design of community-based surveys and intervention studies. Am J Epidemiol. 1999; 149:876–83.
28. Bland JM. Cluster designs: a personal view. http://www.sghms.ac.uk/depts/phs/staff/jmb/clustalk.htm (accessed Nov 2, 2001).
29. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. BMJ. 1998; 317:1171–2.
30. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996; 276:637–9.
31. Elbourne DR, Campbell MK. Extending the CONSORT statement to cluster randomized trials: for discussion. Stat Med. 2001; 20:489–96.

APPENDIX A

The Data Set

center	*resident*	*group*	*yrs*	*delta*	*score1*	*score2*	*pass1*	*pass2*
1	1	0	1	0.8	74.3	75.1	0	1
1	2	0	2	−4.7	75.1	70.4	1	0
1	3	0	2	−2.0	75.5	73.5	1	0
1	4	0	2	5.1	73.6	78.7	0	1
1	5	0	3	−1.1	76.3	75.2	1	1
1	6	0	4	8.3	75.9	84.2	1	1
1	7	0	3	7.4	74.4	81.8	0	1
1	8	0	3	−2.8	75.0	72.2	0	0
2	9	0	1	2.4	74.1	76.5	0	1
2	10	0	1	−4.4	74.2	69.8	0	0
2	11	0	1	−2.0	75.2	73.2	1	0
2	12	0	2	−15.3	76.3	61	1	0
2	13	0	2	−4.7	74.7	70	0	0
2	14	0	2	−0.8	74.2	73.4	0	0
2	15	0	3	5.3	73.6	78.9	0	1
2	16	0	3	−4.0	74.8	70.8	0	0
3	17	1	1	8.2	74.1	82.3	0	1
3	18	1	1	3.1	74.4	77.5	0	1
3	19	1	2	−3.0	74.7	71.7	0	0
3	20	1	2	3.0	74.3	77.3	0	1
3	21	1	2	−0.9	74.2	73.3	0	0
3	22	1	2	−3.3	77.0	73.7	1	0
3	23	1	3	5.5	73.7	79.2	0	1
3	24	1	3	−4.4	74.4	70	0	0
4	25	1	2	9.2	73.3	82.5	0	1
4	26	1	2	3.6	73.9	77.5	0	1
4	27	1	3	3.4	74.9	78.3	0	1
4	28	1	4	6.4	75.6	82	1	1
4	29	1	1	−4.7	76.1	71.4	1	0
4	30	1	3	7.7	76.5	84.2	1	1
4	31	1	4	13.8	73.2	87	0	1
4	32	1	2	9.1	75.0	84.1	1	1
5	33	0	1	−1.2	74.8	73.6	0	0
5	34	0	1	−5.2	74.1	68.9	0	0
5	35	0	1	−3.9	75.9	72	1	0
5	36	0	2	1.9	75.3	77.2	1	1
5	37	0	2	−1.9	75.6	73.7	1	0
5	38	0	2	8.6	74.1	82.7	0	1
5	39	0	3	5.0	75.3	80.3	1	1
5	40	0	3	−3.9	74.6	70.7	0	0
6	41	1	4	8.9	74.1	83	0	1
6	42	1	2	2.4	74.9	77.3	0	1
6	43	1	3	4.9	75.8	80.7	1	1
6	44	1	1	−6.8	74.3	67.5	0	0
6	45	1	3	3.1	74.4	77.5	0	1
6	46	1	3	5.6	75.3	80.9	1	1
6	47	1	4	11.3	75.1	86.4	1	1
6	48	1	2	2.2	76.1	78.3	1	1