The Practice of Informatics

Review Paper

Design and Analysis of Controlled Trials in Naturally Clustered Environments:

Implications for Medical Informatics

JEN-HSIANG CHUANG, MD, MS, GEORGE HRIPTSAK, MD, MS, DANIEL F. HEITJAN, PHD

Abstract In medical informatics research, study questions frequently involve individuals who are grouped into clusters. For example, an intervention may be aimed at a clinician (who treats a cluster of patients) with the intention of improving the health of individual patients. Correlation among individuals within a cluster can lead to incorrect estimates of the sample size required to detect an effect and inappropriate estimates of the confidence intervals and the statistical significance of the intervention effects. Contamination, which is the spread of the effect of an intervention or control treatment to the opposite group, often occurs between individuals within clusters. It leads to an attenuation of the effect of the intervention and reduced power to detect a difference. If individuals are randomized in a clinical trial (individual-randomized trial), then correlation must be taken into account in the analysis, and the sample size may need to be increased to compensate for contamination. Randomizing clusters rather than individuals (cluster-randomized trial) can eliminate contamination and may be preferred for logistical reasons. Cluster-randomized trials are generally less efficient than individual-randomized trials, so the tradeoffs must be assessed. Correlation must be taken into account in the analysis and in the sample-size calculations for cluster-randomized trials.


Medical informatics interventions that are designed to improve the care of patients are frequently aimed not directly at patients but at clinicians who care for multiple patients. For example, a clinical decision support system may give advice to clinicians, yet the important outcome is the patients’ health. When such interventions are studied, determining the appropriate randomization method, sample size, and approach to analysis can be challenging. Even when the intervention is aimed at the individual, natural clustering into clinicians’ practices, families, outpatient
practices, hospital wards, schools, and communities can affect the results. Clustering may be nested\(^1\); for example, patients may be clustered by physicians and physicians may be clustered by multi-physician practices.

In individual-randomized trials, a researcher assigns individuals to a control or intervention group without regard to clustering, or at least in such a way that control patients and intervention patients may be mixed within a cluster. Thus, a given clinician may treat control patients and intervention patients, receiving computer-generated advice on some patients but not on others. If the researcher randomizes individuals but ignores the natural clustering of patients in the analysis, erroneous conclusions may result because of correlation among patients within clusters. Furthermore, contamination due to indirect benefit from the intervention to patients in control groups may affect the results.

In cluster-randomized trials, investigators randomize intact units such as clinicians, families, practices, or other clusters into intervention or control groups. For example, in a study by McDonald et al.\(^2\) 27 practice teams were randomly allocated to use a clinical decision support system or to serve as controls. In another example, McDowell et al.\(^3\) randomized 822 families to receive no influenza vaccine reminder (control group) or to receive a reminder from their clinician, by telephone, or by letter. If the researcher randomizes by cluster, then the power of the study to detect differences between groups will decrease and, if the analysis is not done correctly, erroneous conclusions may be drawn.

A review of evaluations of clinical decision support systems\(^4\) revealed that 24 of 61 studies randomized clusters of patients. Of those 24, only one mentioned sample size calculations, and it did not account for clustering, and only 14 accounted for clustering in the analysis. In this paper, we review the issues surrounding the clustering of patients, covering randomization, sample size, and approaches to analysis.

**Addressing the Pitfalls of Naturally Clustered Environments**

In the discussion that follows, we refer frequently to an example in which the patients are subjects and are clustered by virtue of seeing the same clinician or attending the same clinic. The discussion applies more broadly, however, to students in classrooms, clinicians in practices, hospitals in networks, and others. The clustering of subjects into groups can lead to two major effects—correlation and contamination.

**Correlation**

Two variables are said to be correlated when they change together, in either the same or opposite directions.\(^3\) In a naturally clustered environment, correlation can occur when patients in a cluster have outcomes influenced by some common factor. In most medical informatics studies, correlation will be positive: Patients within groups will tend to have similar outcomes.

Correlation within clusters has several sources.\(^6\)–\(^8\) First, patients often self-select; that is, they choose the clusters to which they belong. For example, patient characteristics such as age, gender, ethnic group, location, or insurance plan may influence their choice of clinicians, leading to similarities across the patients who see a given clinician.

Second, important cluster-level attributes may affect all cluster members in the same way. For example, the rates of clinicians’ compliance with the reminders generated by a clinical decision support system may vary systematically across clinicians. Thus, patients who receive care from the same clinician may be more likely to receive similar treatment than are patients who receive care from different clinicians.

Third, individuals within a cluster often influence each other. For example, the transmission of attitudes and behavior among a cluster of clinicians can lead to all members of the cluster displaying similar behavior. Thus, all clinicians who work on a specific ward or at a specific hospital may have a similar rate of compliance with reminder messages.

Correlation within clusters may affect the results of individual-randomized trials and cluster-randomized trials differently. In individual-randomized trials, most clusters have both intervention and control subjects. Assuming positive correlation, subjects within clusters will look more like each other than subjects in other clusters. Ideally, one would like to compare intervention subjects with control subjects within each cluster (because subjects are most like each other within clusters) and then average the effect across clusters.\(^9\)

If, instead, clustering is ignored, then all the intervention subjects are pooled and all the control subjects are pooled and the two groups are compared. The variance due to differences between clusters (which is relatively large) is mixed with variation.
between subjects within clusters (which is relatively small because of the positive correlation), resulting in unnecessarily large standard deviations, larger \( p \) values, wider confidence intervals, and false-negative results.\(^9\) By accounting for clustering, the researcher can remove a significant source of variance (the clusters) and create more precise estimates.

In cluster-randomized trials, failure to account for clustering may even lead to false-positive results with erroneously small \( p \) values. In this case, each cluster has either intervention subjects or control subjects but not both. Comparing intervention with control subjects therefore requires comparisons across clusters, and the variance due to differences between clusters contributes to the variance of the estimates (rather than being something to be factored out).

Ignoring clustering in the analysis will mix cluster variance (relatively large) with variance between subjects within clusters (relatively small because of the positive correlation) and lead to an underestimate of the overall variance, with inappropriately small \( p \) values and narrow confidence intervals. (For example, if the correlation is high, then adding patients to existing clusters will add little information to the data set, yet an analysis that ignores clustering will count each new patient as an independent observation and exaggerate the precision of its estimates.) A proper analysis that accounts for clustering will correctly estimate the variance of the estimates. In either form of randomization, proper analysis can account for correlation.

### Contamination

Contamination is the spread of the effect of an intervention from one group to another. Contamination occurs when the control group members are exposed to the experimental intervention or the intervention group members are exposed to the control treatment.\(^10\) Contamination is a concern in medical informatics studies. For example, in a study of clinical decision support systems with patients as the randomization factor, clinicians may gain knowledge from use of the system in intervention patients and apply it to control patients.\(^11\) Similarly, in a trial for the prevention of coronary heart disease, participants in the control group may learn about the experimental intervention and adopt it themselves because intervention and control participants are in close proximity and information may be shared.\(^10\)

Contamination leads to the attenuation of the treatment effect, because the control group and the intervention group look more alike. The result is a reduced ability to detect an effect (false-negative result). The researcher can compensate for contamination by increasing the sample size; the effect will be underestimated, but it may still be significantly different statistically from no effect.

The researcher can eliminate the contamination by using a cluster-randomized design, thus separating the control subjects from the intervention subjects. McDonald et al. used a cluster-randomized design to reduce contamination: by randomizing practice teams, the investigators could ensure that the clinicians caring for the control subjects would be isolated from the effects of experience with the clinical decision support system.\(^2\) Similarly, in the study by McDowell et al.,\(^3\) use of families as the randomization unit reduced potential contamination. If an individual were randomized to serve as a control, but the spouse was randomized to the intervention group, then a reminder issued to the spouse might have an influence on the vaccine-seeking behavior of the control subject.

### Randomization Method: Making a Choice

Neither an individual-randomized design nor a cluster-randomized design is superior in all circumstances, so the researcher must make a choice. Assuming that the data will be analyzed properly to account for natural clustering (see subsequent sections), the choice depends on efficiency, contamination, and practical considerations such as ethics, cost, and feasibility.\(^10\)

If contamination is low, then individual-randomized trials tend to be more efficient than cluster-randomized trials and thus tend to require fewer subjects. If individuals are randomized, then each cluster will usually have individuals both in the control and in the intervention groups, so that control and intervention subjects are in effect matched by cluster, making it easy to remove the effect of inter-cluster variability. If clusters are randomized, then each cluster has only control subjects or only intervention subjects. Inter-cluster variability can still be estimated and accounted for, but the analysis is less efficient.

If contamination is likely to be significant, then a cluster-randomized design that eliminates contamination can improve the probability of detecting an effect, and this improvement may be greater than the relative loss of efficiency due to correlation within clusters. In terms of total sample size, cluster-randomized trials may be more efficient than individual-randomized trials when contamination is greater than 30 percent.\(^12\)
Generally, individual-randomized trials should be considered first if an intervention is delivered to individual patients directly. Both correlation within clusters (patients being treated by the same clinician) and contamination are still possible, but as long as clusters are accounted for in the analysis, individual randomization is likely to be more efficient.

If an intervention is delivered to clinicians, a cluster-randomized trial may be superior because contamination is difficult to quantify and may in fact be very high. McDonald et al. argued, however, that some interventions that are aimed at clinicians but that involve complex calculations may be immune to contamination because the clinicians cannot apply the experience to the control subjects without the help of the information system. The justification for the use of the clusters as the units of allocation should be reported explicitly.

Logistical or ethical issues may favor a cluster-randomized design. For example, women invited to participate in a breast cancer screening program may want to discuss their options and the associated risks and benefits with their clinician before making a decision. If only half of a particular clinician’s patients have been invited to participate, patients not invited may feel resentment. For this reason, many trials of cancer screening programs have adopted cluster-randomized designs.

**Design and Analysis of Individual Randomized Trials**

**Handling Correlation in the Analysis of Individual-randomized Trials**

In this section, we consider the two main approaches to accounting for correlation within clusters—fixed effects models and random effects models. Both types of models separate the variability between clusters from the effect of the intervention, thus improving the precision of the estimate of the treatment effect.

In a fixed effects model, the effects of different clusters on the outcome are regarded as fixed but unknown quantities to be estimated. This model assumes that the clusters observed in the experiment are the only clusters of interest. If subjects are clustered by academic department, for example, and if all the departments in an institution (or all the departments of interest) are included in the study, then department can be treated as a fixed effect. Fixed effects models estimate regression coefficients (using linear regression or logistic regression, for example) to embody the effect of each cluster. Strictly speaking, conclusions drawn from the experiment apply specifically to those clusters that were included in the study, although it can be inferred that similar results might be obtained in similar clusters.

Given a well-defined question, fixed effects models usually lead to more precise results, and such models are the only realistic option if there are very few clusters. True random sampling of clusters is rare in clinical trials.

Random effects models assume that the clusters are a random sample drawn from a larger population of potential clusters. For example, a study may include a limited number of clinical practices that constitute a random sample drawn from some larger number of practices. The results of the random effects analysis apply not just to the clusters that were observed but also to the larger population from which they were drawn. The results are thus more generalizable than those for the fixed effects models. Such models offer a more realistic representation of the true uncertainties due to the sampling of clusters, and this is represented by wider confidence intervals for the results. The choice between fixed or random effects models may depend on the sampling situation and on the intended scope of the inferences.

In practice, a given model may have many terms, some of which are random and some of which are fixed. For example, the cluster term may be a random effect, but the intervention term, which takes on only two values (“intervention” or “control”), will be fixed. Such models may be referred to as “mixed” effects models or simply as random effects models. The important issue in this discussion is whether the cluster term is a fixed or random effect.

To illustrate the two models and the consequence of not accounting for clustering, we simulated a data set for an individual-randomized trial in a naturally clustered environment. In the simulation, 100 patients were randomized to treatment (placed on an antihypertension guideline) or to control (no guideline). Patients attended one of ten clinics, and correlation due to self-selection within clinics was expected. Randomization did not take clinics into account. The data set contained four variables—patient age (a continuous variable), clinic identifier (a nominal variable), treatment group (a dichotomous variable), and systolic blood pressure (the continuous outcome variable). The intraclass correlation coefficient for systolic blood pressure within clinics was 0.470 after adjustments for age and for treatment group; thus
there was a significant positive correlation in the data set. We used linear regression (GLM procedure in SAS) for the fixed effects models and linear mixed models (MIXED procedure in SAS) for the random effects model.\textsuperscript{21,22} (The data sets and SAS code are available at http://www.dmi.columbia.edu/home pages/chuangj/cluster/.)

The first model (Table 1) ignored clustering (the clinic variable) in the analysis, resulting in a wide confidence interval for the treatment effect and a non-significant \( p \) value (\( p = 0.136 \)). The second model accounted for clustered as a fixed effect, resulting in a narrower confidence interval and a statistically significant \( p \) value (\( p = 0.013 \)). The third model accounted for clustering as a random effect, resulting in a confidence interval that was only slightly wider than that for the corresponding fixed effect model and a similar \( p \) value (\( p = 0.015 \)). Clearly, the decision to account for clustering (by any method) was most important. The choice between a fixed effects model or a random effects model to account for clustering led to only a subtle difference in this example.

When the outcome variable is binary rather than continuous, a number of approaches may be taken to analyzing individual-randomized trials in naturally clustered environments. The tutorial by Agresti and Hartzel, for example, compared Mantel-Haenszel methods (FREQ procedure in SAS), logistic regression (GENMOD procedure in SAS), and a generalized linear mixed model (NLMIXED procedure in SAS).\textsuperscript{19}

Compensating for Contamination in Individual-randomized Trials

To make up for the attenuation of the treatment effect due to contamination, the researcher may increase the sample size by the factor

\[
\frac{1}{(1 - \text{contamination})^2}
\]

, where “contamination” is the proportion of the treatment effect that is attenuated.\textsuperscript{10,23} For example, if the treatment effect is reduced by 10 percent because of contamination, then the sample size should be increased by approximately 25 percent. The degree of attenuation is difficult to estimate, however, and the calculation of adjusted sample sizes is often impractical in medical informatics studies. In these cases, a cluster-randomized design may be better.

**Table 1**

Analysis of Naturally Clustered Data When Individuals Are Randomized

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Fixed Effects</th>
<th>Random Effects</th>
<th>( p ) Value of Treatment Difference</th>
<th>Treatment Difference (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fixed effects model that ignores clustering (incorrect model)</td>
<td>Age, treatment</td>
<td>-</td>
<td>0.136</td>
<td>4.5 mmHg (–1.4–10.4)</td>
</tr>
<tr>
<td>2</td>
<td>Fixed effects model that accounts for clustering</td>
<td>Age, treatment, clinic</td>
<td>-</td>
<td>0.013</td>
<td>6.0 mmHg (1.3–10.7)</td>
</tr>
<tr>
<td>3</td>
<td>Random effects model that accounts for clustering</td>
<td>Age, treatment</td>
<td>Clinic</td>
<td>0.015</td>
<td>5.8 mmHg (1.1–10.5)</td>
</tr>
</tbody>
</table>

**Design and Analysis of Cluster-randomized Trials**

**Choice of Study Design**

When researchers randomize clusters rather than individuals, several common designs are available—completely randomized, stratified, and matched-pair.\textsuperscript{14} In a completely randomized cluster study, each cluster is assigned with a predefined probability to one of the possible intervention groups, and assignments are made independently of each other. In a stratified randomization cluster study, clusters are presorted into strata according to characteristics that are likely to be associated with the outcome, and then clusters within strata are randomly assigned to intervention groups. Completely randomized and stratified designs may also employ a blocking strategy to ensure that the number of clusters per treatment group is approximately balanced.\textsuperscript{23–25} In a matched-pair cluster design, pairs of clusters are matched on the basis of characteristics that are likely to be associated with the outcome, and then one randomly chosen cluster in each pair is assigned to the intervention group and the other is assigned to the control group.

Matched-pair and stratified designs can reduce the potential for baseline imbalance (imbalance in factors that are likely to be correlated with outcomes), especially in a small study.\textsuperscript{26} Just as in individual-ran-
randomized trials, stratification and matching should be done only on cluster-level variables that are known to be highly correlated with outcome.\textsuperscript{14,27} Frequently used stratification factors include cluster size, geographic area, socioeconomic indicators, and characteristics of clinicians and hospitals.\textsuperscript{28–31}

**Sample Size Estimation in Cluster-randomized Trials**

Applying standard sample size formulas to cluster-randomized trials may lead to underestimation of the required sample size. For completely randomized cluster design, we multiply standard sample size estimates by a design effect (\(Deff\)) term given by

\[
Deff = 1 + (m - 1)\rho
\]

where \(m\) denotes the estimated average cluster size (average number of individuals per cluster, assuming that all clusters are of a similar size) and \(\rho\) denotes the intracluster correlation coefficient. Methods for sample-size estimation for stratified and matched-pair designs are presented elsewhere.\textsuperscript{14,32,33}

The intracluster correlation coefficient represents the ratio of between-cluster variability to total variability, where \(0 \leq \rho \leq 1\):

\[
\rho = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_w}
\]

where \(\sigma^2_b\) denotes the between-cluster variance component and \(\sigma^2_w\) denotes the within-cluster variance component.\textsuperscript{24} If \(\rho\) equals 0, then individuals within the same cluster are no more likely to be correlated with each other than with individuals in different clusters. If \(\rho\) equals 1, there is no variability within a cluster, and individuals within the same cluster respond identically.\textsuperscript{23,27} The correlation coefficient can be obtained from previously published studies or from the analysis of pilot data. A random effects model (for example, MIXED procedure in SAS) can supply the estimates of the variance components (\(\sigma^2_b\) and \(\sigma^2_w\)).

An example is taken from Kerry and Bland,\textsuperscript{34} in which a behavioral intervention trial was proposed to reduce smoking rates among patients seen by primary care practice groups. The clusters were patients seen at individual practices, and the researchers used completely randomized clusters. The goal was to estimate the sample size needed to achieve 90 percent power at the two-sided 5 percent level of significance for detecting a 5 percent prevalence difference in smoking. The smoking rate in the control group was 22 percent and the between-practice variance (\(\sigma^2_p\)) of the smoking rate was 0.0014, which was obtained from an earlier study.\textsuperscript{35}

Based on a standard sample size equation for an individual-randomized trial, the total number of patients required per intervention group was 1,318. The mean of the smoking rates (\(\bar{p}\)) between the control (22 percent) and intervention (17 percent) groups was 19.5 percent. The value of within-practice variance (\(\sigma^2_w\)) was 0.1570, which was estimated from \(\bar{p}(1 - \bar{p})\). Based on \(\sigma^2_b\) and \(\sigma^2_w\), the intracluster correlation coefficient was 0.0088.

If the number of patients per practice, \(m\), was 50, then the design effect was 1.43, and the total number of patients required was 1,900, implying 38 practices in each intervention arm (rather than 27 practices in each arm if the original sample size of 1,318 had been used). Failing to account for clustering might have led to a negative result due to insufficient sample size.

**Handling Correlation in the Analysis of Cluster-randomized Trials**

To avoid false-positive or otherwise misleading conclusions, it is critical to account for correlation within clusters in cluster-randomized trials. There are two overall approaches to the analysis—cluster-level analyses and individual-level analyses that account for clustering. (Note that the level of the randomization may be different from the level of analysis. One may randomize clusters but still analyze by individuals as long as this is done correctly.)

Cluster-level analyses aggregate the individual observations using cluster means, proportions, or log odds, resulting in a single value per cluster. Standard statistical methods are then applied. In general, cluster-level analyses can be used for completely randomized, matched-pair, and stratified cluster designs. For example, a standard or a weighted two-sample Student \(t\) test or a Wilcoxon rank-sum test (a nonparametric approach also known as the Mann-Whitney \(U\) test) can be used to cluster means\textsuperscript{36,37} in a completely randomized design. The weighted \(t\) test is preferred to the standard \(t\) test if cluster sizes vary significantly. A paired \(t\) test, a weighted paired \(t\) test, or a signed rank test can be applied at the cluster level in matched-pair and in stratified designs. Although cluster-level analyses address the problem of correlation within clusters, individual-level covariates cannot be analyzed because the data have been aggregated.

Individual-level analyses preserve the individual
observations but still account for the correlation within clusters. Random effects models and generalized estimating equations (GEEs) are commonly used. Individual-level analyses can be applied to any of the randomized cluster designs. They account for both individual- and cluster-level variation and thus may produce more precise estimates than cluster-level analyses. In addition, individual-level analyses can be applied to nonrandomized designs because they support the control of baseline differences between groups; nevertheless, randomized studies are preferred when they are possible.

The choice between individual- and cluster-level analyses in cluster-randomized trials depends on several factors. In the absence of adjustment for individual-level covariates (that is, if the individual level information is ignored, other than its contribution to the cluster mean), the two approaches will produce similar results. When the focus of the inference is at the cluster level—for example, clinicians’ compliance with clinical guidelines—cluster-level analyses are preferred because the units of intervention are the same as the units of evaluation. Furthermore, when the number of clusters per intervention group is small, it may be better to use a cluster-level analysis. Individual-level analyses that account for clustering may be more appropriate when the individual outcomes (for example, the status of diabetic control) are the primary focus. In circumstances where individual covariates need to be adjusted, individual-level analysis should be used.

According to our previous review, 11 of 14 clinical decision support system studies that did account for clustering in their analyses used cluster-level analyses. They most frequently used the two-sample t test and analysis of variance. The other three studies used individual-level analyses. Of these, one used a random effects model and two used generalized estimating equations.

We used data from Kerry and Bland to illustrate the analysis of cluster-randomized studies. They studied the effect of a clinical guideline on the appropriateness of radiology referrals in 34 general medical practices (Table 2). Intervention practices received a copy of the guideline and control practices did not. Practices (clusters) were of widely varying size. The outcome was dichotomous—whether or not the patient’s care conformed to the guideline. We used the SAS System for all analyses. The results are shown in Table 3.

### Table 2
Cluster-randomized Trial Data

<table>
<thead>
<tr>
<th>Intervention Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Conforming Requests/Total</td>
<td>No. of Conforming Requests/Total</td>
</tr>
<tr>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>20/20</td>
<td>7/7</td>
</tr>
<tr>
<td>7/7</td>
<td>33/37</td>
</tr>
<tr>
<td>15/16</td>
<td>32/38</td>
</tr>
<tr>
<td>28/31</td>
<td>23/28</td>
</tr>
<tr>
<td>18/20</td>
<td>16/20</td>
</tr>
<tr>
<td>21/24</td>
<td>15/19</td>
</tr>
<tr>
<td>6/7</td>
<td>7/9</td>
</tr>
<tr>
<td>5/6</td>
<td>19/25</td>
</tr>
<tr>
<td>25/30</td>
<td>90/120</td>
</tr>
<tr>
<td>53/66</td>
<td>64/88</td>
</tr>
<tr>
<td>4/5</td>
<td>15/22</td>
</tr>
<tr>
<td>33/43</td>
<td>52/76</td>
</tr>
<tr>
<td>32/43</td>
<td>14/21</td>
</tr>
<tr>
<td>16/23</td>
<td>83/126</td>
</tr>
<tr>
<td>44/64</td>
<td>14/22</td>
</tr>
<tr>
<td>4/6</td>
<td>21/34</td>
</tr>
<tr>
<td>10/18</td>
<td>56/76</td>
</tr>
<tr>
<td>Mean (SD) –</td>
<td>81.5 (12.0)</td>
</tr>
<tr>
<td>TOTAL*</td>
<td>341/429</td>
</tr>
<tr>
<td>%</td>
<td>79.5</td>
</tr>
</tbody>
</table>


*The percentage conforming based on the raw total does not equal the mean of the percentage conforming for each cluster, because clusters are of unequal size.

- **Ignoring clustering in the analysis (incorrect approach).** If clustering is ignored, a simple 2x2 contingency table is formed and a chi-square test (FREQ procedure in SAS) can be used to compare the two proportions. The difference in conformance (7 percent) appeared to be highly significant ($p=0.008$). This naïve approach must not be used in practice, because it will count every individual as an independent observation despite positive correlation within clusters. The result will be artificially precise estimates and artificially low $p$ values.

- **Two-sample t test (cluster-level analysis).** For a simple cluster-level analysis, we used a two-sample t test (TTEST procedure in SAS) with the percentage of patients conforming per practice as the observations. The mean difference in conformance (8 percent) was not significant ($p=0.074$).

- **Weighted t test (cluster-level analysis).** The number
of patients per practice varies widely in this example, so a weighted $t$ test may be preferred. We used a $t$ test weighted by the number of patients per cluster (TTEST procedure with WEIGHT statement in SAS). The weighted mean difference (7 percent) was statistically significant ($p = 0.044$).

- Random effects model (individual-level analysis). We used the SAS GLIMMIX macro to implement a generalized linear mixed model. In this example, the random effects variable was the practice (cluster); the other variables were fixed effects. The odds ratio (1.5) was significantly different from 1 ($p = 0.046$). (With Version 7 of SAS, the NLMIXED procedure can be used to fit a similar model.)

- Generalized estimating equations (individual-level analysis). Generalized estimating equations have become important tools for analyzing longitudinal and other correlated data. They generate estimates similar to those from standard logistic regression but adjust for the effect of clustering. The structure of the correlation can be estimated from the data. We used the GENMOD procedure with REPEATED statement in SAS to implement generalized estimating equations. The odds ratio (1.5) was significantly different from 1 ($p = 0.040$).

In this example, the naïve approach gave a deceptively low $p$ value—far lower than any other approach tried. A cluster level analysis using a simple $t$ test was not statistically significant due to widely varying cluster size. All other approaches had similar results, with $p$ values ranging from 0.040 to 0.046. Therefore, the single most important decision is to account for clustering in the analysis. Failure to do so can lead to incorrect results. The differences among the various methods of accounting for clustering can be subtle, although a poor choice can result in some loss of power.

**Conclusions**

In many medical informatics studies, individuals are naturally grouped into clusters. Because of correlation within clusters and contamination between intervention groups within clusters, false-positive or false-negative results may be obtained if clustering is not accounted for in sample size estimates and in the analysis. Cluster-randomized designs can eliminate contamination and may sometimes be appropriate for logistical reasons, but they must be analyzed appropriately.

The authors thank Lyn Dupré for her editing.

**References**


---

**Table 3**

<table>
<thead>
<tr>
<th>Statistical Test</th>
<th>Test Statistic (df)</th>
<th>$p$ Value</th>
<th>Effect Size</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignoring clustering in the analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square test</td>
<td>$\chi^2 = 6.95$ (1 df)</td>
<td>0.008</td>
<td>PD = 7%</td>
<td>2%–12%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR = 1.5</td>
<td>1.1–2.0</td>
</tr>
<tr>
<td>Cluster-level analyses:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-sample $t$ test</td>
<td>$t = 1.84$ (32 df)</td>
<td>0.074</td>
<td>MD = 8%</td>
<td>-1%–17%</td>
</tr>
<tr>
<td>Weighted $t$ test</td>
<td>$t = 2.09$ (32 df)</td>
<td>0.044</td>
<td>WMD = 7%</td>
<td>0%–14%</td>
</tr>
<tr>
<td>Individual-level analyses that</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>account for clustering:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random effects model</td>
<td>$t = 2.00$ (1,097 df)</td>
<td>0.046</td>
<td>OR = 1.5</td>
<td>1.0–2.3</td>
</tr>
<tr>
<td>Generalized estimating equations</td>
<td>$z = 2.06$</td>
<td>0.040</td>
<td>OR = 1.5</td>
<td>1.0–2.1</td>
</tr>
</tbody>
</table>

*ABBREVIATIONS: PD indicates proportion difference; MD, mean difference; WMD, weighted mean difference; OR, odds ratio; CI, confidence interval; df, degree of freedom.*