Community Dent Oral Epidemiol 2006; 34: 241–51

Community Dentistry and Oral Epidemiology

Methods

Multilevel analysis of group-randomized trials with binary outcomes

Hae-Young Kim[1], John S. Preisser[2], R. Gary Rozier[3] and Jayasanker V. Valiyaparambil[3]

[1]Dental Research Institute, School of Dentistry, Seoul National University, Seoul, Korea, [2]Department of Biostatistics, School of Public Health University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [3]Department of Health Policy and Administration, School of Public Health University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Kim H-Y, Preisser JS, Rozier RG, Valiyaparambil JV. Multilevel analysis of group-randomized trials with binary outcomes. Community Dent Oral Epidemiol 2006; 34: 241–51. © Blackwell Munksgaard, 2006

Abstract – *Objectives:* Many dental studies have assessed the effectiveness of community- or group-based interventions such as community water fluoridation. These cluster trials, of which group-randomized trials (GRTs) are one type, have design and analysis considerations not found in studies with randomization of treatments to individuals (randomized controlled trials, RCTs). The purpose of this paper is to review analytic methods used for the analysis of binary outcomes from cluster trials and to illustrate these concepts and analytical methods using a school-based GRT. *Methods:* We examine characteristics of GRTs including intra-class correlation (ICC), their most distinctive feature, and review analytical methods for GRTs including group-level analysis, adjusted chi-square test and multivariable analysis (mixed effect models and generalized estimating equations) for correlated binary data. We consider two- and three-level modeling of data from a cross-sectional cluster design. We apply the concepts reviewed using a GRT designed to determine the effect of incentives on response rates in a school-based dental study. We compare the results of analyses using methods for correlated binary data with those from traditional methods that do not account for ICC.
Results: Application of traditional analytic methods to the dental GRT used as an example for this paper led to a substantial overstatement of the effectiveness of the intervention. *Conclusions:* Ignoring the ICC among members of the same group in the analysis of public health intervention studies can lead to erroneous conclusions where groups are the unit of assignment. Special consideration is needed in the analysis of data from these cluster trials. Randomization of treatments to groups also should receive more consideration in the design of cluster trials in dental public health.

Key words: cluster trials; dental survey incentives; group-randomized trials; response rates; school-based survey research

R. Gary Rozier, Department of Health Policy and Administration, The University of North Carolina at Chapel Hill 1105F McGavran-Greenberg Hall, CB No. 7411 Chapel Hill, NC 27599–7400, USA
Tel: 919 966 7388
Fax: 919 966 6961
e-mail: gary_rozier@unc.edu

Submitted 13 June 2005;
accepted 28 March 2006

Cluster trials evaluate interventions delivered to intact social groups or clusters, while outcomes are measured on individual members of those groups. Murray et al. (1) recommend the group-randomized trial (GRT), which is a specific type of cluster trial, as the gold standard when allocation of identifiable groups is undertaken, just as the randomized controlled trial (RCT) is the gold standard when the study design requires allocation of individuals. The GRT is widely used in public health evaluations. Varnell et al. (2) found 60 such studies published in just two public health journals during a 5-year period. Considerable recent research has also been devoted to the development, evaluation and synthesis of procedures for their design and analysis (3, 4).

Like other areas of public health, dentistry has a strong tradition in the evaluation of community and practice interventions. Many studies have been undertaken to assess the effectiveness of public health interventions such as community water fluoridation (5), school-based sealant programs (6) and oral health education programs (7). These types of interventions, by their very design, target entire communities or groups, so random assignment of treatment conditions at the level of

individuals usually is not possible or appropriate. Often these trials are designed to test the effects of patient clinical interventions, but they are implemented on a clinic- or practitioner-level for reasons such as avoiding contamination of treatment effects (8).

Only a few of the cluster trials reported in the dental public health literature have used randomization in the assignment of groups to experimental conditions, and most of these have appeared in the last few years (9–15). Those that do use random assignment in the implementation of a cluster trial often do not appropriately apply GRT design and analytic methods nor do they provide sufficient details about the analysis used to obtain the results. The potential for correlation of clinical measurements taken within the same mouth to cause a type I error has been recognized for many years (16). Randomizing groups and analyzing individuals while ignoring intra-class correlation is a mistake akin to dental studies randomizing individuals but collecting and analyzing data at the surface- or site-level without regard for the intra-mouth correlation of the data.

The purpose of this paper is to review analytic methods used for the analysis of GRTs. The paper considers two- and three-level analysis of binary outcomes of data from studies using cross-sectional cluster designs. We illustrate the concepts and analytical methods included in our review using data from a school-based study in which students nested within classrooms nested within schools are treated as three levels with interventions assigned to schools. The concepts discussed in the paper also apply to group trials without randomization, a more common design in dental public health studies.

A primer on group-randomized trials

The major characteristics of the GRT and RCT are contrasted in Table 1. Unlike RCTs where individuals are assigned, groups are the units for random assignment of interventions in GRTs. However, like the RCT, the outcomes of interest are measured at the individual level.

The most important difference between GRTs and RCTs is the presence of intra-class correlation (ICC) resulting from the similarity of values for the outcome measures taken from members of the same group. The correlation arises because individuals within the same group tend to be more alike than do members across different groups. These similarities arise because group members share the same environment and interact with each other.

Table 1. Comparative characteristics of the group-randomized trial (GRT) and the randomized controlled trial (RCT)

	GRT	RCT
Unit of assignment	Group	Individual
Unit of observation	Individual	Individual
Relationship of subjects	Correlation within group	Independent[a]
Adjustment of sample size for design effect	Needed	Not needed

[a]Exceptions such as multiple or repeated measurements in an individual exist.

The impact of the ICC in GRTs is akin to that of the ICC in surveys where the use of cluster sampling methods can result in measurements of interest with larger variances and less precise estimates than would result from identification of individuals through simple random sampling. The degree of extra variation is measured by the 'design effect', which is the ratio of the variance of an estimator of the intervention effect in a GRT design to that of the corresponding estimator from a RCT design. The design effect is sometimes called the 'variance inflation factor' because, as in studies employing complex sample surveys where the latter term receives common usage, the extra variation of a GRT relative to a RCT is evident when the value of the design effect exceeds one. The design effect should be considered in the design and analysis of GRTs if one wants to achieve the same precision as would be expected in an RCT with randomization of individuals (17). Generally, ignoring extra variation due to intra-cluster correlation leads to an underestimation of the variance of the intervention effect, which, in turn, leads to inflation of type I error, or the chances of rejecting a true null hypothesis (18, 19).

The design effect for a one-stage sample of clusters is defined as $\phi = 1 + (m-1)\rho$, where m denotes the size of each group or cluster. If group sizes vary, m is replaced with an 'adjusted' mean group size defined below. The ICC, measuring the magnitude of correlation among group members, is commonly represented as $\rho = \mathrm{corr}(y_{ij}, y_{ij'}) = \frac{\sigma^2}{\sigma^2 + \sigma_e^2}$, where i, and j denote groups and subjects in the groups, respectively; where $j \neq j'$; σ^2 and σ_e^2 denote between-group

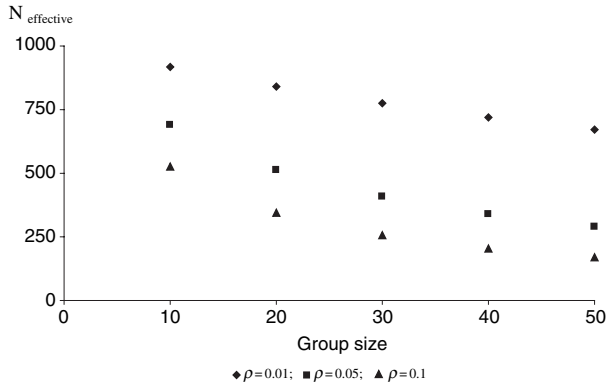

Fig. 1. The relationship between effective sample size ($N_{effective}$) and group size by various degrees of intra-class correlation (ICC or ρ): The sample size used in group-randomized trials (N_{GRT}) is fixed at 1000.

variance and within-group variance, respectively. This formulation assumes a common ICC across groups. A small within-group variance relative to between-group variance will make ρ, and, in turn, ϕ, large. The design effect is useful in relating the sample size needed in a GRT (N_{GRT}) to the effective sample size ($N_{effective}$), or the sample size needed for a comparative RCT. The N_{GRT} can be calculated easily by using the following formula: $N_{GRT} = N_{effective} \times \phi$ (20).

A common characteristic of GRTs is the use of a small number of groups with a relatively large number of subjects within each group. In this case, even a small ICC can result in a substantial design effect (3, 4). Figure 1 depicts the relationship between group size (m) and $N_{effective}$, and demonstrates that even a small ICC (ρ) of 0.01 could result in an effective sample size ($N_{effective}$) of only about 675 subjects from a total of 1000 subjects in a GRT (N_{GRT}) as the mean group size approaches 50.

Many estimators of ICC have been proposed for binary outcomes. Zou and Donner (21) give closed-form formulae for three different ICC estimators and their variances including the widely used ANOVA estimator. They recommend the kappa type estimator ($\hat{\rho}_{FC}$) of Fleiss and Cuzick that performs well when there are 50 or more clusters.

Statistical analysis of GRTs

The statistical analysis of data from a GRT requires a different strategy than does an RCT. A comparison of analytic methods that can be used in the two types of studies with binary outcomes is presented in Table 2 and discussed in the following paragraphs.

Group-level analysis

A common approach for analysis of data from a GRT is to consider the groups as the units of analysis and compare the proportion of subjects in each group with the particular outcome of interest (3, 4, 22). In this case, the sample size is the number of groups, and one can simply use a *t*-test or ANOVA to compare the proportions (12). Data often do not satisfy the assumption of a normal distribution with equal variance, especially if group sizes are varied. In that situation, nonparametric procedures, such as the Wilcoxon rank sum test or Kruskal–Wallis test are more useful than parametric tests.

Adjusted chi-square test for individual-level analysis

Similar to the standard Pearson chi-square test for independent samples, one can use an adjusted chi-square test to compare the event rates in correlated data derived from a GRT (4, 22–24). If the intervention has three arms (C = 3), as does the example used in this paper, then the adjusted chi-square statistic is

$$\chi^2_A = \sum_{k=1}^{3} \frac{M_k\left(\hat{P}_k - \hat{P}\right)^2}{\phi_k \hat{P}(1-\hat{P})}$$

where intervention k = 1,2,3; M_k is the total number of individuals receiving intervention k; m_{ki} is the number of individuals in ith group receiving intervention k; $\hat{P}_k$ is the event rate in intervention k; $\hat{P}$ is the overall event rate in the total sample; ϕ_k is the design effect for condition k,

Table 2. Analysis methods for the group-randomized trial (GRT) and the randomized controlled trial (RCT) with binary outcomes

	GRT	RCT
Group-level analysis	Comparison using proportions	N/A
Test at individual level	Adjusted chi-square test	Chi-square test
Multivariable modeling	Generalized linear mixed modeling (GLMM) or generalized estimating equation (GEE)	Single-level logistic model

$\phi_k = 1 + (\bar{m}_{Ak} - 1)\hat{\rho}$, where the adjusted group size is

$$\bar{m}_{Ak} = \sum_{i=1}^{n_k} m_{ki}^2 / M_k.$$

Note that if all group sizes within the kth condition are equal then $\bar{m}_{Ak}$ reduces to the common group size, m.

Multivariable regression analysis of correlated binary data

Two general classes of models are applicable for the multivariable analysis of correlated binary data: (i) generalized linear mixed models (GLMM), also called conditional or cluster-specific models, which are commonly estimated by maximum likelihood; and (ii) marginal or population-averaged models, estimated by generalized estimating equations (GEE) (1, 4, 22, 25–28). While different link functions are possible (29), we present specific forms based on the logit link, which is generally applied for binary outcomes.

Mixed effect models

The GLMM based on the logit link is called the logistic random effects model or the logistic-normal model because of the common assumption that random effects are normally distributed. For a two-level model there is a single random effect, u_i, for each group $i = 1,\ldots,K$. The two-level logistic-normal model is

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = X_{ij}'\beta + u_i.$$

Note that $p_{ij} = E(Y_{ij} \mid u_i)$ is the probability that the jth subject in the ith group (e.g. school) is an event case conditional upon the value of the random effect, u_i, for the ith group. In mixed effects models, β represents within group change (i.e. conditional on u_i) (30, 31). Specifically, we assume $u_i \sim N(0, \sigma_b^2)$,the common assumption for the two-level or random intercept logistic-normal model (31). Assuming normality for the random effects completes the full-likelihood specification of the distribution of correlated binary responses across groups. The ICC is not modeled directly, but rather heterogeneity of groups is modeled explicitly via the random effect variance, σ_b^2, which is the between-group scaled variance (i.e. on the logit scale). This is the only variance component in the model as the residual error is the theoretical variance for the binary outcome. To estimate ICC we can estimate σ_e^2 by $\hat{P}(1 - \hat{P})$. Next, noting that the unscaled variance σ^2 has the approximate relationship $\sigma^2 \approx \sigma_b^2[\hat{P}(1 - \hat{P})]^2$, provides a quick formula for ICC (3, p. 234–240):

$$\rho = \mathrm{corr}(y_{ij}, y_{ij'}) = \frac{\sigma^2}{\sigma^2 + \sigma_e^2} \approx \frac{\sigma_b^2}{\sigma_b^2 + 1/\hat{P}(1 - \hat{P})}.$$

Now suppose there are two levels of nesting. The three-level logistic-normal model is

$$\log\left(\frac{p_{ijs}}{1 - p_{ijs}}\right) = X_{ijs}'\beta + u_i + u_{is},$$

where $p_{ijs} = E(Y_{ijs} \mid u_i, u_{is})$ is the probability that the jth subject (e.g. student) in the sth subcluster (e.g. classroom) from the ith cluster or group (e.g. school) is an event case conditional upon the value of the random effects. We assume the subcluster and cluster random effects are statistically independent, $u_i \sim N(0, \sigma_b^2)$, and $u_{is} \sim N(0, \sigma_s^2)$.

Estimation of β in GLMMs is computationally complex due to the need to handle the random effects (32).

Generalized estimating equations

The population-averaged logistic model is expressed by

$$\log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right) = X_{ij}'\beta^*,$$

where $\mu_{ij} = E(Y_{ij})$ is the marginal probability that the jth subject in the ith group is an event case. The parameter β^* represents the population-averaged change (30, 32) and is generally not equal to the cluster-specific parameter β (33). Unlike the mixed effects model, the GEE approach does not explicitly account for group-to-group heterogeneity. The variance function for binary data, $\mathrm{var}(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$, and a working correlation structure to account for ICC, completes specification of the model. For the cross-sectional design, an exchangeable correlation structure is often assumed, i.e. $\mathrm{corr}(y_{ij}, y_{ij'}) = \rho$. Estimation of β^* is performed by iteratively reweighted least squares (34). A model-based variance estimator provides valid inference, in a large sample sense, when the correlation structure has been correctly specified; the empirical or 'robust' variance estimator is valid even if the correlation structure is misspecified (34). However, inflated type I error is common when GEE is implemented with empirical variance estimators for GRTs with fewer than 40 groups (27, 35).

A bias-corrected variance estimator improves the accuracy of inference with GEE when the number of groups is small (36). GEE estimators of ρ tend to underestimate ICC when there are a small number of clusters and so they are not recommended in most cluster trials situations.

An example: improving response rates in a school-based dental study

Overview of incentive study

Data used for our example of analytical principles for GRTs are derived from a study to assess the effects of teacher and student monetary incentives on parent–child unit response rates to a school-based survey with a clinical examination of the child and a self-completed questionnaire for the parent. The rationale for the study is that response rates in population-based oral health surveys have declined in recent years, and little empirical evidence is available to guide public health practitioners in planning and conducting school-based surveys in ways that achieve maximum subject participation (37). Parents and students were assigned to one of three intervention types (teacher incentive = \$20; student incentive = \$1, \$2, \$5 for 1st, 6th, 10th grade, respectively; and no monetary incentive). In addition to the monetary incentives, parents of all participating students received results of the oral examination. The clinical examination of the child required positive consent by a parent.

Because this work was undertaken as a pilot to a planned larger survey of schoolchildren and their parents in North Carolina a cost-saving design was used at the potential loss of some statistical power. The study employed a stratified, cluster-RCT with posttest data collection only and where incentive conditions were randomly assigned to three county units within each of two strata defined by county income level (median per capita income >\$40 000 vs. <\$30 000). In the high income strata, county units were equivalent to counties and within those counties, three schools, one of each grade level (1st, 6th, 10th grade), were recruited and assigned the same intervention; there was one exception as two grade levels from one school participated. The low income strata also consisted of three county units, but the three schools for one of these units were selected from three different counties. Thus, schools from eight counties participated in the study, but the county unit was the unit of randomization. The final sample included 1482 students from 70 classrooms in 17 schools in six county units. The number of students per school varied from a minimum of 60 to a maximum of 160 with a mean of 87.1.

In the terminology of GRTs, this study employed a four-level, multilevel design with county units as clusters, schools as sub-clusters, classrooms as sub-sub-clusters and student-parent pairs as units of observation. We used this design for a number of scientific and practical reasons. We felt that randomization of counties and assignment of schools located within them to the same intervention as a unit would help limit contamination among groups that could result from teachers and administrations within the same school system having contact with each other. Any knowledge among teachers or students about the incentive structure could affect the internal validity of the study by affecting their behaviors in ways such as 'resentful demoralization' or 'compensatory rivalry' (38). Assignment of multiple schools within counties also made the trial more efficient because training in trial implementation would require less time away from usual responsibilities for the six county-based dental hygienists who delivered the intervention. Each of them needed to become familiar with only one intervention. We also believed that administrators within a single school system would be more likely to approve the trial if it involved the same intervention for their county.

For the purposes of this paper, we treat the data as if they came from a three-level nested cross-sectional GRT with posttest outcome only (3). Ignoring county units (level 4) as a source of variation, students (observations, level 1) are nested within classrooms (sub-clusters, level 2) within schools clusters (level 3). While the actual randomization strategy employed provides no basis for ignoring the potential source of between county unit variation associated with randomization at the county unit level, there is some empirical motivation for doing so as efforts to model county unit as an additional random effect (beyond the three-level structure or selected two-level structures) were unsuccessful; in other words, the variance component associated with county unit was estimated to be zero when added to the multilevel models described in the next section. Even so, our main reason for ignoring county units is to provide illustrative analyses of three-level data (if hypothetical) and not to provide a definitive analysis of the incentives study. A more costly but definitive follow-up GRT to the

pilot with adequate power to address study questions would recruit schools from different counties and randomize interventions at the school level. Therefore, analyses presented are those that would be applicable to such a study. Conclusions as they pertain to the actual pilot study should be viewed with caution.

Outcome and ICC estimate

The main interest of this study was to determine whether parent–child units respond differently according to three different types and amounts of monetary incentives. The outcome is a binary variable, whether parent–child units respond or not to both the questionnaire and clinical oral examination. We obtained an unadjusted ICC estimator of 0.07 using the Fleiss-Cuzick estimator and 95% confidence interval, 0.01–0.32, by the method of inverting a modified Wald test (21). The design effect is computed as $\phi = 1 + (87.1-1)\times 0.07 = 7.02$. Similarly, we obtained a model-adjusted ICC estimate of 0.075 from the two-level GLMM model applied below.

Application of analysis strategies to incentive study

Group-level analysis using proportions

The overall unadjusted response rates were 0.35, 0.43 and 0.34 for the teacher incentive, student incentive and no incentive groups, respectively. Differences in response rates were not significant ($P = 0.5417$) using the Kruskal–Wallis test applied to school-level proportions.

Individual-level analysis using adjusted chi-square test

We compared the unadjusted and adjusted standard Pearson chi-square tests to evaluate the effect of incentives on response rates at the individual level. The difference between the two resulting statistics was large. The adjusted two-degrees of freedom chi-square statistic and corresponding *P*-value ($\chi^2_2 = 1.37$; $P = 0.50$) indicated lack of a statistically significant difference in contrast to a significant difference obtained from the unadjusted statistic ($\chi^2_2 = 12.86$; $P = 0.002$) in favor of student incentives. The statistical significance of the unadjusted chi-square test may be attributed to inflated type I error resulting from a failure to adjust for ICC.

Multivariable analysis

Initially, we define the two-level logistic-normal model to explain the response rate, p_{ij}, of the jth student, in the ith school:

$$\begin{aligned}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \beta_0 + \beta_{11}(\text{incentive} = \text{teacher}) \\ &+ \beta_{12}(\text{incentive} = \text{student}) \\ &+ \beta_{21}(\text{income} = \text{high}) + \beta_{31}(\text{grade} = 1) \\ &+ \beta_{32}(\text{grade}=6) + \beta_{41}[(\text{income} = \text{high}) \\ &\times (\text{grade} = 1)] + \beta_{42}[(\text{income} = \text{high}) \\ &\times (\text{grade} = 6)] + u_i\end{aligned}$$

where, $u_i \sim N(0, \sigma^2_b)$. Thus β_0 represents the baseline log odds of response associated with the control incentive condition for schools with the 10th grade from the strata of low income counties; β_{11} and β_{12} are the cluster-specific log odds ratios of response (relative to the control condition) for teacher and student incentives, respectively, adjusting for income and grade. In view of the stratification of the study design, we retained income, grade, and their interactions in the model, as well as the two incentive dummy variables, regardless of their statistical significance.

To estimate parameters of the GLMM for our example, we used the SAS GLIMMIX procedure, which provides approximate maximum likelihood estimates (39). We also used PROC GLIMMIX to fit the three-level model that treats classrooms as sub-clusters through specification of a second random effect. Finally, we used the SAS GENMOD procedure with repeated statement to fit the population-averaged model with the same covariates by GEE. For GEE, we used a working exchangeable correlation structure and bias-corrected standard errors for the estimated parameters because the empirical sandwich standard errors or model-based errors may be comparatively unstable for GRTs with relatively small numbers of 17 groups (40).

Results from single-level logistic regression models are compared with those of the two two-level GEE models considering schools or classrooms as the second level (Table 3). Relative to the multilevel models, the standard errors from the single-level logistic regression are much smaller and lead to an erroneous finding of a statistically significant effect due to student incentives. In contrast, neither the GEE nor mixed model approaches in Table 4 identified a statistically significant incentive effect. From the two-level GEE model adjusting for schools, the estimated population averaged odds ratio of response to student incentive relative to control was $e^{0.260} = 1.29$. No effect of teacher incentive was found.

Table 3. Comparison of a naïve (incorrect) single-level logistic regression model and two-level logistic regression models fitted with GEE adjusting for either the classroom level or school level, predicting probability of response

	Logistic		GEE			
Level(s)	Single		Two			
Adjustment	None		Classroom		School	
	β (SE)	*P*-value	β (SE)[a]	*P*-value	β (SE)[b]	*P*-value
Incentive (teacher)	0.0315 (0.1349)	0.8156	0.0837 (0.2078)	0.6870	−0.0512 (0.4577)	0.9124
Incentive (student)	0.4230 (0.1361)	0.0019	0.4334 (0.2406)	0.0717	0.2601 (0.4216)	0.5418
Income (high)	0.1126 (0.2029)	0.5790	0.0645 (0.4068)	0.8740	0.0490 (0.6751)	0.9442
Grade (1st)	0.7332 (0.2011)	0.0003	0.7646 (0.3941)	0.0523	0.8122 (0.6676)	0.2225
Grade (6th)	0.8296 (0.2048)	<.0001	0.8711 (0.4326)	0.0440	0.7670 (0.5439)	0.1586
Income (high) × Grade (1st)	0.1175 (0.2737)	0.6677	0.1511 (0.4862)	0.7561	0.0610 (0.8916)	0.9442
Income (high) × Grade (6th)	−1.0222 (0.2764)	0.0002	−1.0653 (0.5281)	0.0437	−0.9354 (0.9549)	0.3174
Intercept	−1.0630 (0.1722)	<.0001	−1.0998 (0.3252)	0.0007	−0.9712 (0.4758)	0.0444
$\hat{\rho}$			0. 1029		0.0191	

[a]Robust standard errors shown.
[b]Bias-corrected standard errors shown.

Table 4. Two-level GLMM and three-level GLMM models predicting probability of response

Levels	Two				Three	
Adjustment	Classroom		School		Both	
	β (SE)	*P*-value	β (SE)	*P*-value	β (SE)	*P*-value
Incentive (teacher)	−0.015 (0.2721)	0.9663	0.0670 (0.3893)	0.8633	0.0138 (0.4119)	0.9733
Incentive (student)	0.4015 (0.2681)	0.1345	0.4934 (0.3704)	0.1830	0.4337 (0.3968)	0.2746
Income (high)	0.1180 (0.3949)	0.7651	0.1213 (0.5282)	0.8184	0.1186 (0.5709)	0.8355
Grade (1st)	1.1229 (0.3833)	0.0035	1.1138 (0.4956)	0.0248	1.2898 (0.5385)	0.0167
Grade (6th)	0.5758 (0.3915)	0.1415	0.5693 (0.4959)	0.2511	0.5011 (0.5422)	0.3555
Income (high) × Grade (1st)	−0.1414 (0.5288)	0.7892	−0.2360 (0.7190)	0.7428	−0.3032 (0.7704)	0.6940
Income (high) × Grade (6th)	−0.7582 (0.5343)	0.1561	−0.7495 (0.7196)	0.2978	−0.6571 (0.7730)	0.3955
Intercept	−1.1785 (0.3305)	0.0007	−1.1353 (0.4373)	0.0267	−1.2023 (0.4731)	0.0293
$\hat{\sigma}^2_b$ (classroom)	0.5756 (0.1615)	0.0004			0.4292 (0.1380)	0.0018
$\hat{\sigma}^2_s$ (school)			0.3532 (0.1991)	0.0768	0.2923 (0.2351)	0.3150

Results from two-level GLMM models adjusting schools or classrooms as the second level, and three-level GLMM models regarding both school-levels and classroom-levels are presented in Table 4. The estimated school-specific odds ratios of response to student incentive relative to control was $e^{0.434} = 1.54$ for the three-level GLMMs. Generally, the two-level mixed models produced overall results similar to the three-level model in terms of statistical significance, albeit with a tendency to produce smaller standard errors. This similarity is particularly true for the two-level model that specifies classroom as cluster and ignores potential intraclass correlation at the school level. As in the GEE analyses, no effect of teacher incentive was found in any of the GLMMs.

The estimated variance of the cluster-specific random effects was $\hat{\sigma}^2 = 0.35$ for the two-level logistic-normal model giving an estimated ICC of 0.075. In contrast, the estimated exchangeable correlation (ICC) was $\hat{\rho} = 0.019$ for the model based on GEE. Parameters from logistic-normal models and those from GEE have the approximate relationship $\beta_{\text{Logistic}-\text{Normal}} \approx \beta^*_{\text{GEE}}\sqrt{1 + 0.35\sigma^2}$ (30, 31). The estimates in Tables 3 and 4 reflect this relationship with the mixed model parameter estimates generally being greater in absolute value than the GEE estimates for most variables.

Sample size calculation for future study

Assuming an ICC of 0.07 obtained from this pilot study, and applying the usual sample size formulae for a two-sided test of binary outcomes in a GRT (i.e. equation (6) of reference (41), a two-condition GRT comparing student incentives (with assumed response rate 0.43) to no incentives (response rate 0.34) with randomization of incentives to schools (instead of county-units), and

enrolling 87 students per school, requires 37 clusters (schools) per condition to achieve a power of 0.8 with type I error $\alpha = 0.05$.

Discussion

A comparison of results from the single- and multilevel regression models presented in this paper demonstrated that the use of a standard logit model in the analysis of GRTs can underestimate the uncertainty of estimates of intervention effects and lead to faulty conclusions. While the effects of student incentives ($P = 0.002$) on parent–child participation was highly statistically significant in the single-level model, it was not significant in other multilevel models. This finding supports the conclusion that single-level ordinary logistic regression analysis is not appropriate for the analysis of the correlated binary outcomes from GRTs.

While it is clear that single-level ordinary logistic regression is not appropriate for binary data from GRTs, the better choice of a modeling strategy between GEE for marginal models and the conditional GLMM model is less clear. These alternative modeling approaches have received a lot of attention in research, particularly the interpretation of model parameters and the performance of statistical procedures for their estimation. Because both approaches account for the correlation of outcomes within a group, the choice of which approach to use primarily depends on consideration of the characteristics of a particular trial.

The conditional model is appealing because it explicitly models the between-group heterogeneity via group-level random effects, while marginal models consider it as a nuisance factor. For a trial using a linear model for a continuous outcome variable, interpretation of the parameter estimates is the same in both methods; however for non-identity link models (i.e. nonlinear models) such as the logistic model for binary data, both the relative magnitude of parameters as well as their respective interpretations are generally different. If one is interested in group differences for binary outcomes, the marginal model is more relevant because it gives parameter estimates related directly to the average response. The conditional model is used when the intervention effect on a particular community is desired (18, 31) because these models provide estimates of the effectiveness of the intervention that is conditional on the value of the group-specific random effects. This community-specific effect is not directly observable in the data, as each community is only observed under one intervention condition, but rather it is only obtained through distributional assumptions about the random effects. On the basis of the statistical properties (such as control for type I error) of the estimation procedures when the number of clusters is small, the GLMM model has generally been preferred over GEE for GRTs (18). Many software packages implement the methods discussed here for GRTs with a binary outcome including HLM, several SAS procedures including GLIMMIX and NLMIXED, MIXOR, MLwiN, SPSS, Stata (42) and SUDAAN (43).

With respect to fitting GLMMs to three-level binary data, when possible it is best to fit three-level models as opposed to two-level models because the latter approach may introduce bias at some level (32, 44). However, given occasional difficulties in fitting three-level mixed models to data with two levels of clustering, researchers sometimes ignore one level of clustering and adjust only for the other cluster factor. In an application of a smoking prevention and cessation project where conditions were randomized at the school level and interventions delivered at the classroom level, standard errors of regression coefficients for the two-level models (either ignoring classroom or school as cluster factors) tended to be smaller than standard errors from the three-level model (32). Their result is consistent with the mixed model results for our dental study presented here and suggests that the statistical significance of fixed effects may be overstated with a simpler two-level model.

Among two-level models applied to three-level data, a simulation study motivated by developmental toxicity data sets (e.g. multiple outcomes nested within pups nested with litters) suggests that the second-level mixed model (e.g. if applied to a GRT it would specify a random effect for classroom, but ignore school as a cluster factor) may also overstate statistical significance of third-level fixed effects (e.g. condition factor in a GRT) (44). On the other hand, the same study showed that the third-level mixed model (e.g. a two-level model that specifies a random effect at the third level for schools, but ignores the second-level cluster factor of classrooms) is preferable for assessing interventions in a GRT but may perform worse for assessing the statistical significance of first-level (e.g. student) and second-level (e.g. classroom) factors.

Although these simulations were not designed with large cluster sizes and small numbers of clusters, the primary characteristics of GRTs, the results suggest preference for the third-level over the second-level mixed model for GRTs when a three-level model can not be fit, with the caveat that secondary findings regarding any first- or second-level effects may be biased. We note the third-level model is appropriate for the dental study in this regard because only third-level effects were collected and analyzed.

The effect of the ICC estimate of 0.07 on sample size requirements was substantial compared to what would be required for an RCT. While it is common to use the unadjusted ICC values for estimation of sample size requirements, some investigators have reported around a 50% reduction in ICC estimates by adjustment for individual- and cluster-level variables along with baseline values for the outcome variable (45, 46). Generally, a good strategy is to use relevant stratification in order to address imbalances in outcomes across strata and to obtain a smaller design effect. Generally, the problem of estimating ICCs from GRTs is a difficult one as precision is often poor because of the small number of clusters upon which the estimate is based.

In planning a GRT, the anticipated value for ICC usually is, by necessity, based on the most relevant published studies. Yet, attention to appropriate analysis and reporting of GRTs in the public health literature is less than optimal (47), and we have few examples of ICC values derived from evaluations of community-based interventions in dentistry. Therefore, it is very important not only to consider the appropriateness of the research design and its analysis, as is the case with all research, but to make ICC estimates available to other researchers considering implementation of cluster studies.

Some recent publications have documented ICC estimates for several nondental outcomes in various contexts (46, 48). Campbell et al. (49) suggested three dimensions should be considered when reporting an ICC – a description of the dataset, including characteristics of the outcome and the intervention; information on how the ICC was calculated; and information on the precision of the ICC. It also is important that investigators report other factors related to the design of GRTs, such as number of clusters, their average size, and specific estimation procedures used in the analysis (50).

Conclusion

Many studies have been implemented in dental health research to assess the effectiveness of a community- or group-based intervention such as community water fluoridation or school-based dental health programs. Most of these studies do not apply the appropriate principles for the design and analysis of cluster trials. The most important feature of cluster trials is the existence of ICC caused by correlation among members of the same group. Ignoring the ICC can lead to false conclusions. Special consideration is needed in designing cluster trials and analyzing data from them. Further, groups rarely are randomized to interventions in these trials. The randomized group trial (GRT), as the alternative gold standard for the RCT, should be emphasized in dental public health research.

Acknowledgments

Data used for illustrative purposes in this paper were collected as part of a research project entitled: 'Effectiveness of community-wide strategies to promote oral health' funded by the Centers for Disease Control and Prevention, Grant Number U48/CCU415769.

References

1. Murray DM, Varnell SP, Blistein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health 2004;94:423–32.
2. Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: a review of recent practices. Am J Public Health 2004;94:393–9.
3. Murray DM. Design and analysis of group-randomized trials. New York, USA: Oxford University Press; 1998.
4. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. New York, USA: Oxford University; 2000.
5. McDonagh M, Whiting P, Bradley M, Cooper J, Sutton A, Chestnutt I, et al. A systemic review of public water fluoridation. York, UK: NHS Centre for Reviews and Dissemination, University of York; 2000.
6. Truman BI, Gooch BF, Sulemana I, Gift HC, Horowitz AM, Evans CA, et al. The Task Force on Community Preventive Services. Reviews of evidence on interventions to prevent dental caries, oral and pharyngeal cancers, and sports-related craniofacial injuries. Am J Prev Med 2002;23:21–54.
7. Kay E, Locker D. A systematic review of the effectiveness of health promotion aimed at improv-

ing oral health. Community Dent Health 1998;15: 132–44.
8. Jamtvedt G, Young JM, Kristoffersen DT, Thomson O'Brien MA, Oxman AD. Audit and feedback: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2003;3:CD000259.
9. Masouredis CM, Hilton JF, Grady D, Gee L, Chesney M, Hengl L, et al. A spit tobacco cessation intervention for college athletes: three-month results. Adv Dent Res 1997;11:354–9.
10. Gordon JS, Andrews JA, Lichtenstein E, Severson HH, Akers L. Disseminating a smokeless tobacco cessation intervention model to dental hygienists: a randomized comparison of personalized instruction and self-study methods. Health Psychol 2005;24:447–55.
11. Bahrami M, Deery C, Clarkson JE, Pitts NB, Johnston M, Rickettes I, et al. Effectiveness of strategies to disseminate and implement clinical guidelines for the management of impacted and unerupted third molars in primary dental care, a cluster randomized controlled trial. Br Dent J 2004;197:691–6.
12. Kiang H, Bian Z, Tai BJ, Du MQ, Peng B. The effect of a bi-annual professional application of APF foam on dental caries increment in primary teeth: 24-month clinical trial. J Dent Res 2005;84:265–8.
13. Gansky SA, Ellison JA, Rudy D, Bergert N, Letendre MA, Nelson L, et al. Cluster-randomized controlled trial of an athletic trainer-directed spit (smokeless) tobacco intervention for collegiate baseball athletes: results after 1 year. J Athl Train 2005;40:76–87.
14. Finch C, Braham R, McIntosh A, McCory P, Wolfe R. Should football players wear custom fitted mouthguards? Results from a group randomised controlled trial. Inj Prev 2005;11:242–6.
15. Blinkhorn AS, Gratrix D, Holloway PJ, Wainwright-Stringer YM. A cluster randomised, controlled trial of the value of dental health educators in general dental practice. Br Dent J 2003;195:395–400.
16. Chilton NW. Design and analysis in dental oral research. Philadelphia: J.B. Lippincott Company, p. 309–10.
17. Cornfield J. Randomization by group: a formal analysis. Am J Epidemiol 1978;108:100–2.
18. Feng Z, Diehr P, Petersen A, McLerran D. Selected statistical issues in group randomized trials. Annu Rev Public Health 2001;22:167–87.
19. Goldstein H. Multilevel statistical models. Second Edition. London, England: Edward Arnold; 1995.
20. Snijder T, Bosker R. Multilevel analysis. London: Sage Publications Ltd.; 1999.
21. Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 2004;60:807–11.
22. Preisser JS. Cluster trials. Encyclopedia Biopharmaceut Stat, 2005;1. Available at: http://www.dekker.com/sdek/abstract. Last accessed 01 May 2006.
23. Begg MD. Analyzing k (2×2) tables under cluster sampling. Biometrics 1999;55:302–7.
24. Panageas KS, Begg MD, Grbic JT, Lamster IB. Analysis of multiple 2×2 tables with site-specific periodontal data. J Dent Res 2003;82:514–7.
25. Sashegyi AI, Brown KS, Farrel PJ. Application of a generalized random effects regression model for cluster-correlated longitudinal data to a school-based smoking prevention trial. Am J Epidemiol 2000;152: 1192–200.
26. Kuss O. How to use SAS® for LOGISTIC regression with correlated data. SUGI 27. 2002. Paper 261–27. Available at: http://www2.sas.com/proceedings/sugi27/p261–27.pdf. Accessed January 2006.
27. Bellamy SL, Gibberd R, Hancock L, Howley P, Kennedy B, Klar N, et al. Analysis of dichotomous outcome data for community intervention studies. Stat Methods Med Res 2000;9:135–59.
28. Diggle PJ, Heagerty PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data, 2nd edition. Oxford: Oxford University Press; 2002.
29. McCullagh P, Nelder JA. Generalized linear models, 2nd ed. London: Chapman and Hall; 1989.
30. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics 1988;44:1049–60.
31. Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject specific approaches for analyzing repeated binary outcome. Am J Epidemiol 1998;147:694–703.
32. Gibbons RD, Hedeker D. Random effects probit and logistic regression models for three-level data. Biometrics 1997;53:1527–37.
33. Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. Stat Methods Med Res 1992;1:249–73.
34. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
35. Murray DM, Hannan PJ, Wolfinger RD, Baker WL, Dwyer JH. Analysis of data from group-randomized trials with repeated observations on the same groups. Stat Med 1998;17:1581–600.
36. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics 2001;57:126–34.
37. Macfarlane TV, Worthington HV. Some aspect of data analysis in dentistry. Community Dent Health 1999;16:216–9.
38. Trochim WMK. The research methods knowledge base, 2nd edition. Cincinnati: Atomic Dog Publishing, 2001.
39. SAS Inc. The GLIMMIX procedure, 2005. Available at: http://support.sas.com/rnd/app/papers/glimmix.pdf. Accessed January 2006.
40. Hammill BG, Preisser JS. A SAS/IML software program for GEE and regression diagnostics. Comput Stat Data Anal 2005;in press.
41. Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. Stat Med 2003;22:1235–54.
42. Stata Corporation. Stata statistical software: Release 8.0. College Station, TX: Stata Press; 2003.
43. Research Triangle Institute. SUDAAN language manual, Release 9.0. Research Triangle Park, NC: Research Triangle Institute; 2004.
44. Ten Have TR, Kunselman AR, Tran L. A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. Stat Med 1999;18:947–60.
45. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. Eval Rev 2003;27:79–103.

46. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J Clin Epidemiol 2004;57:785–94.
47. Smith PJ, Moffatt MEK, Gelskey SC, Hudson S, Kaita K. Are community health interventions evaluated appropriately? A review of six journals. J Clin Epidemiol 1997;50:137–46.
48. Murray DM, Phillips GA, Birnbaum AS, Lytle LA. Intraclass correlation for measures from a middle school nutrition intervention study: estimates, correlates, and applications. Health Educ Behav 2001;28:666–79.
49. Campbell MK, Grimshaw JM, Elbourne DR. Intracluster correlation coefficients in cluster randomized trials: empirical insights into how they should be reported. BMC Med Res Methodol 2004;4:9.
50. Guo G, Zhao H. Multilevel modeling for binary data. Annu Rev Sociol 2000;26:441–62.