Published in final edited form as:
Epidemiology. 2007 September ; 18(5): 552–560.

The Importance and Role of Intracluster Correlations in Planning Cluster Trials

John S. Preisser[*], **Beth A. Reboussin**[†], **Eun-Young Song**[‡], and **Mark Wolfson**[‡]

[*]*Department of Biostatistics, University of North Carolina School of Public Health, Chapel Hill, North Carolina*

[†]*Department of Biostatistical Sciences, Division of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina*

[‡]*Department of Social Sciences and Health Policy, Division of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina*

Abstract

There is increasing recognition of the critical role of intracluster correlations of health behavior outcomes in cluster intervention trials. This study examines the estimation, reporting, and use of intracluster correlations in planning cluster trials. We use an estimating equations approach to estimate the intracluster correlations corresponding to the multiple-time-point nested cross-sectional design. Sample size formulae incorporating 2 types of intracluster correlations are examined for the purpose of planning future trials. The traditional intracluster correlation is the correlation among individuals within the same community at a specific time point. A second type is the correlation among individuals within the same community at different time points. For a "time × condition" analysis of a pretest–posttest nested cross-sectional trial design, we show that statistical power considerations based upon a posttest-only design generally are not an adequate substitute for sample size calculations that incorporate both types of intracluster correlations. Estimation, reporting, and use of intracluster correlations are illustrated for several dichotomous measures related to underage drinking collected as part of a large nonrandomized trial to enforce underage drinking laws in the United States from 1998 to 2004.

Cluster trials evaluate interventions delivered to intact social groups or clusters, such as communities, churches, schools, workplaces, and medical practices, whereas outcomes are measured on members of those groups.[1,2] The distinctive feature of cluster trials is the presence of intracluster (or intraclass) correlation among members within groups that arises from restricting assignment of the interventions to groups instead of to individuals. Failure to account for the intracluster correlation within clusters will likely lead to 2 shortcomings: an underpowered study and inflated Type I error rate of hypothesis tests relating to the intervention.[1,2] Proper planning of cluster trials is based upon sample size formulae that use hypothesized intracluster correlation values, often based on estimates from earlier trials.[3-6]

Although sample size formulae based upon the intracluster correlation for a posttest-only cluster trial design are fairly well established, many cluster trials have multiple time points for which more than one type of intracluster correlation arises. In the case of a pretest–posttest nested cross-sectional design, characterized by different groups of individuals within clusters sampled at 2 time points, 2 intracluster correlations may be defined. The traditional intracluster

Correspondence: John S. Preisser, Department of Biostatistics, CB # 7420 School of Public Health, University of North Carolina, Chapel Hill, NC 27599−7420. E-mail: jpreisse@bios.unc.edu.

correlation is the correlation among individuals within the same community at a specific time point. A second intracluster correlation is the correlation among individuals within the same community at different points in time. Extensions of the posttest-only sample size formulae are needed to refine sample size determination for multiple time point designs.

The sample size formula adopted in planning a cluster trial depends upon the statistical analysis to be used and, more specifically, the test statistic for the null hypothesis of “no intervention effect.” The proposed analysis will depend upon the trial's experimental design, including whether the intervention condition is randomized to groups. For example, in a randomized pretest–posttest nested cross-sectional design, the trial planners are concerned with how the statistical analysis should handle the pretest data to maximize statistical power for the test of intervention, given that a sufficiently large number of clusters are enrolled to ensure baseline balance among covariates and outcome. In short, the focus is on the comparison of posttest means. In contrast, potential bias from baseline imbalance is an additional concern in planning a nonrandomized cluster trial. In this case, a “time × condition” analysis may be appropriate because its test statistic is the difference in change in the mean outcome over time between intervention and control conditions. If the pretest intervention and control condition means are equal, a time × condition analysis will tend to have lower statistical power than a posttest only analysis because the former uses a test statistic that is a function of 4 means, resulting in greater variance than a test statistic that is a function of only 2 (posttest) means. In sum, the manner of accounting for pretest information in the analysis of a randomized cluster trial has implications for statistical power, whereas the choice may affect both power and bias for nonrandomized trials.

This article discusses the estimation of intracluster correlations and particularly their use in sample size formulae in the design of cluster trials. Although statistical analysis and power have received attention for modeling change in Gaussian outcomes,[7,8] insufficient consideration has been given to binary outcomes.[9,10] Focusing on large cluster trials with binary outcomes, a generalized estimating equations (GEE) approach is applied for the estimation of the 2 types of intracluster correlation.[11] In planning a nonrandomized cluster trial with a pretest–posttest nested cross-sectional design, the paper discusses a sample size formula for a GEE time × condition analysis that incorporates the 2 types of intracluster correlations.[12] Finally, the paper reports intracluster correlation estimates and estimates of their precision for several dichotomous measures of underage drinking from a large nonrandomized cluster trial to enforce underage drinking laws.[12]

METHODS

The Enforcing Underage Drinking Laws Program

The Enforcing Underage Drinking Laws Program, launched by the United States Office of Juvenile Justice and Delinquency Prevention in 1998, is the largest federal initiative focused on reducing underage drinking in United States history. Each of the 50 states and the District of Columbia received significant funding and technical assistance to support state and local efforts to enforce laws related to alcohol use by underage persons and to prevent underage drinking. A major component of the program involved discretionary grants provided to states on a competitive basis. Selected states awarded subgrants to communities (cities or counties) according to criteria that varied across states. Randomization was not employed, and the evaluation team had no control over the choice of intervention communities. However, control communities with characteristics similar to the intervention communities were selected by identifying those with propensity scores similar to the intervention communities.[12] The propensity score was a scalar summary of several community characteristics captured by federal census and other external sources, and identified a priori as being likely related to the grantee selection process and the major outcomes measured by the youth survey.

The evaluation of the impact of the discretionary grants, or “national evaluation,” was conducted with a nested cross-sectional design.[2,13,14] Data were collected using a variety of surveys; the focus in this work is on 3 annual telephone surveys using distinct samples of between 15 and 20 youths (age 16–20 years) in each selected community. For each community, data collection was conducted preintervention (or early in the intervention period), 1 year later, and 2 years later. A nested cross-sectional design was chosen over a nested cohort design because the long intervention period might result in substantial dropout in a cohort design.[1,14] Also, interest focused on the change in the population over time as opposed to within-individual change.[12]

The national evaluation was conducted using data collected during 3 funding cycles. The first began in 1999 with 52 intervention and 52 control communities in 9 states. The second began in 2000 with 16 intervention and 16 control communities from 7 states. The final cycle began in 2002 with 34 intervention and 34 control communities from 8 states. All but 2 communities participated in exactly one funding cycle. A total of 10,865 observations from 202 communities were used in the analysis of underage drinking outcomes.

Nine dichotomous measures of underage drinking use, alcohol risk behaviors, and negative consequences of alcohol use from the survey of youth were analyzed (Table 1). The following individual and contextual community-level variables were examined as covariates because they may partly explain the magnitude of intraclass correlations for alcohol use behaviors: age, sex, and community-level variables that are possibly characteristic of disadvantaged communities, namely, percentage of households with female head with no husband, percentage foreign born, and median income.

Statistical Analysis

The sample was predominantly white and well-balanced with respect to sex (Table 2). Sixteen and seventeen-year-olds were over-represented in the sample compared with 18, 19, and 20-year-olds. Observed prevalence for each outcome is reported in Table 3. Analyses of the intervention, reported elsewhere,[15] used GEE with the simple “exchangeable” working correlation matrix[9,13,16] to fit time × condition logistic regression models; use of the “robust” empirical covariance estimator provided valid large sample inference even if the correlation structure was misspecified.

The estimation of intracluster correlations in this article employs an extension of GEE that jointly specifies one set of estimating equations for the parameters in the logistic model for the probability that an individual reports the behavior, and a second set of estimating equations to estimate the parameters in the correlation model.[11] In this approach, applied separately for the various behavior outcomes, a correlation model based upon 2 intracluster correlations is specified, the “within-time” correlation between outcomes from different youths at the same time (α_0), and the “between-time” correlation between 2 outcomes from different youths at different times (α_1).[12] The approach produces estimates of the standard error for both intracluster correlations as well as for their covariance. As in the ordinary GEE,[16] the extended GEE approach provides valid inference for assessing intervention effects, assuming the marginal model for the probability of behavior is correctly specified, even if the correlation model is misspecified.

Three sets of logistic models are fit for the probability that a youth reports the behavior. An initial set of models includes the design variables of condition (ie, control versus intervention), time, and funding cycle, as well as their pairwise interactions. A second set of models includes these terms in addition to the individual characteristics of age and sex. A third and final set of models adds both individual and the community characteristics, which serves 2 purposes. First, it may help to address any postsample selection imbalances among covariates in this nonrandomized study. Second, it may partly explain the magnitude of within-cluster correlation.[5] For example, smaller intracluster correlations in the third model, may suggest that intracluster correlations in the first or second models are partly explained by variation in community characteristics.

Sample Size Determination

A general approach to power calculations for cluster trials is based upon an analysis of community summary statistics according to the study design.[12] The general set-up is to specify the hypothesis of interest as $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \mu_1 - \mu_2$ is the intervention effect and $\mu_h = E(S_{hi})$ is the expected value of the summary statistic, S_{hi}, for the i-th community receiving the intervention ($h = 1$) or control ($h = 2$) treatment condition. Specific examples of S_{hi} relevant in the application of GEE to the national evaluation data are discussed below. As communities are assumed to be statistically independent, deriving an expression for the variance of the community summary statistic in each condition (intervention and control) is the critical step in determining sample size. For the ith community receiving the hth condition, let $\sigma_h^2 = \mathrm{Var}(S_{hi})$, and let m be equal to the number of subjects in each community at each time-point. Constant variance within condition results from assuming that m is constant. The number of communities needed per condition (n) to test the intervention using a two-sided test with α significance level and power 1 – β is

$$n=\frac{\left(\sigma_1^2+\sigma_2^2\right)\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{\delta^2} \tag{1}$$

where z_c is the (100× c)th percentile of the standard normal distribution. For small n, Equation 1 may be refined using the t-distribution.[2] Let $\phi(\cdot)$ define the cumulative distribution function of the standard normal distribution, and

$$d=\sum_{i=1}^{n} S_{1i}/n - \sum_{i=1}^{n} S_{2i}/n,$$

an unbiased estimator of δ. Power (1 – β)is

$$1-\beta=\Phi\left(\frac{\delta}{\sqrt{\mathrm{var}\,(d)}}-z_{1-\alpha/2}\right) \tag{2}$$

where $\mathrm{var}(d) = (\sigma_1^2 + \sigma_2^2)/n$.

Of specific interest are sample size formulae for pretest-posttest nested cross-sectional designs as they pertain to binary outcomes. In this design, the total number of individuals sampled per community, or cluster size, is $2m$. Let π_{ht} be the probability of the outcome for an individual at time t ($t = 0$ for pretest, $t = 1$ for posttest) from a community having condition h. Subscript i for the individual is not needed because the probability of outcome is assumed to depend only upon the treatment status and time point.

In randomized cluster trials with a moderate to large number of communities, a planned GEE analysis need not adjust for pretest since groups may be expected to be balanced with respect to outcomes and covariates as a result of randomization. Statistical inference may be based on the posttest-only logistic model

$$\mathrm{logit}\,(\pi_{h1})=\beta_0+\delta_0 x_1 \tag{3}$$

where $x_1 = 1$ if $h = 1$ (intervention) and 0 if $h = 2$ (control); and $\delta_0 = \mu_1 - \mu_2$, where $\mu_1 = \mathrm{logit}(\pi_{11})$ and $\mu_2 = \mathrm{logit}(\pi_{21})$, is the log odds ratio comparing odds of response in the posttest period for subjects in intervention communities to the odds of response for subjects in control communities. Suppose S_{hi} is the logit of the observed proportion reporting the behavior of those sampled at posttest from the *i*th community having condition *h*. Then *d* is an approximately unbiased estimator of δ_0 whose large sample variance depends upon the approximate variance of S_{hi} under $H_1 : \delta_0 \neq 0$:

$$\sigma_h^2 \approx \frac{\phi}{m v_{h1}}, \tag{4}$$

where $\nu_{ht} = \pi_{ht}\,(1 - \pi_{ht})$ and $\phi\ = 1 + (m - 1)\alpha_0$ is the design effect. When $\alpha_0 > 0$, ϕ represents a multiplicative increase on the sample size required in a cluster trial to obtain a given level of power relative to the sample size required in a clinical trial with randomization of individuals. Inserting Equation 4 into Equation 1 and substituting δ_0 for δ gives

$$n = \frac{\phi\,(1/v_{11} + 1/v_{21})\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{m[\operatorname{logit}(\pi_{11}) - \operatorname{logit}(\pi_{21})]^2}, \tag{5}$$

the posttest-only cluster trial design sample size for binary outcomes. Similarly, inserting Equation 4 into Equation 2 gives the power formula. The equivalency of Formula 5 to a general GEE method of sample calculation for the Wald test statistic corresponding to δ_0 in Equation 3 is shown in the Appendix. Alternatively, one can conduct sample size calculations for a planned linear model for the binary outcome (ie, use of identity link in Equation 3) using Equation 6 of Preisser et al.[12]

For a nonrandomized cluster trial with a nested pretest–posttest cross-section design, Formula 5 may be inadequate because adjustment for pretest response is needed in the analysis due to baseline imbalance. While, for the national evaluation, the estimation of intracluster correlations is based upon a model applied to data from 3 time points, sample size considerations for a future nonrandomized cluster trial address the anticipated effect at a single follow-up with respect to baseline. Thus, the comparison of intervention and control communities is operationalized as a one-degree-of-freedom contrast for the difference in the change in expected outcome from pretest to posttest. An appropriate sample size formula targets the contrast

$$\delta_1 = [\operatorname{logit}(\pi_{11}) - \operatorname{logit}(\pi_{10})] - [\operatorname{logit}(\pi_{21}) - \operatorname{logit}(\pi_{20})],$$

the regression coefficient for the time × condition interaction in the logistic regression model for a youth's response at time *t* under the *h*th condition:

$$\operatorname{logit}(\pi_{ht}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta_1\,(x_1 \times x_2) \tag{6}$$

where $x_2 = t$. Specifically, δ_1 is the difference in log odds ratios for pretest/posttest changes between intervention and control conditions. Equivalently, $\exp(\delta_1)$ is the multiplicative factor by which the pretest/posttest odds ratio for communities under the intervention condition exceeds that of the control condition. For a community *i* having condition *h*, define S_{hi} as the logit of observed proportion reporting the behavior at posttest minus the logit of the observed proportion reporting the behavior at pretest. The corresponding mean over all communities within a condition is $\Sigma_t^n = 1\ S_{hi}/n$, an approximately unbiased estimator of $\mu_h = \operatorname{logit}(\pi_{h1}) - \operatorname{logit}(\pi_{h0})$. The approximate variance[12] of S_{hi} is:

$$\sigma_h^2 = \frac{1}{m}\left\{\phi\left(\frac{1}{v_{h1}} + \frac{1}{v_{h0}}\right) - \frac{2m\alpha_1}{\sqrt{v_{h1} v_{h0}}}\right\} \tag{7}$$

Inserting Equation 7 into Equation 1, and setting δ to δ_1 gives the pretest-posttest cluster trial design sample size formula for binary outcomes

$$n = \frac{\left\{\phi \sum_{h=1}^{2} \sum_{t=0}^{1} \left(\frac{1}{v_{ht}}\right) - 2m\alpha_1 \left[\sum_{h=1}^{2} \left(\frac{1}{\sqrt{v_{h1} v_{h0}}}\right)\right]\right\}\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{m\{[\operatorname{logit}(\pi_{11}) - \operatorname{logit}(\pi_{10})] - [\operatorname{logit}(\pi_{21}) - \operatorname{logit}(\pi_{20})]\}^2} \tag{8}$$

Note that the resulting expression depends upon values of the 4 probabilities π_{ht}. Often researchers are unwilling or unable to specify particular values of the π_{ht}. Then values of π_{ht} may be chosen to give conservative values of σ_h^2 in Equation 4 or Equation 7 accordingly. Finally, inserting Equation 7 into Equation 2 and setting δ to δ_1 gives the corresponding formula for power. Alternatively, one can conduct sample size calculations for a planned linear model analysis of the proportions (ie, GEE identity link function) using Equation 9 of Preisser et al[12] and $\delta_1 = (\pi_{11} - \pi_{10}) - (\pi_{21} - \pi_{20})$.

RESULTS

Intracluster Correlation Estimates

Table 3 provides intracluster correlation estimates for several binary outcomes from the national evaluation from the model based upon 3 timepoints and 3 funding cycles. Estimates obtained from the extended GEE approach for the within-time (α_0) and between-time intracluster correlation (α_1) as well as their 95% large sample confidence intervals are reported. The upper confidence bound for each intracluster correlation (eg, for α_0) is given by $\text{UCB}=\widehat{\alpha}_0+1.96 \quad \text{se}\,(\widehat{\alpha}_0)$ and the lower confidence bound is given by LCB = $\widehat{\alpha}_0 - 1.96 \quad \text{se} \quad (\widehat{\alpha}_0)$, where se $(\widehat{\alpha}_0)$ is the standard error based upon the robust variance estimator.[11] The estimated covariance between $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ is also reported for the purpose of conducting power for a range of (α_0, α_1) as illustrated in the next section. Three different sets of intracluster correlations are reported based upon the covariates included in the model for the probability a youth reports a behavior. For each measure, the first row gives the intracluster correlations based upon the model for the probability of the behavior that depends only upon study design variables: time point (year), cycle, intervention versus control condition, and their twoway interactions. The second row gives intracluster correlation estimates based upon a probability model that also adjusts for covariates age and gender. The third row gives intracluster correlations based upon the logistic model that additionally adjusts for the community level variables.

Intracluster correlation estimates range from 0.003 to 0.026 with modest differences in within-time and between-time intracluster correlations. Within-cluster correlations estimates for binge drinking, past 30-day alcohol use and attempt to purchase alcohol are larger than estimates for those outcomes reported by some authors[4,17] but similar in magnitude to those reported by others.[5,18,19] Generally, correlations adjusting for individual and community variables are smaller than design-adjusted correlations or those adjusting for age and gender in addition to design variables. SAS/IML software for applying the extended GEE is available.[20]

Sample Size Results

Consider a future cluster trial to reduce underage drinking. Suppose the primary outcome is past 30-day alcohol use, and a pretest-posttest (2 time-points) nested cross-sectional design is planned with equal numbers of intervention and control communities and $m = 15$ youth to be surveyed from each community at each time point. Assume $\pi_{10} = \pi_{20} = \pi_{21} = 0.40$ and $\pi_{11} = 0.30$ (ie, $\delta_1 = 0.44$) such that a 25% decline in underage drinking is anticipated. Applying Equations 1 and 7, and using the unadjusted intracluster correlation values of $\alpha_0 = 0.0261$ and $\alpha_1 = 0.0219$ from Table 3, results in $n = 48$ communities per condition needed to provide approximately 80% power to detect the desired effect based upon a 2-sided test at the 0.05 significance level.

A sensitivity analysis of power considers a range of values for α_0 and α_1 using information in Table 3. Because combinations of the 2 types of intracluster correlation are required, the univariate confidence intervals reported in Table 3 have limited utility, since the joint 95% confidence region is not rectangular, but rather is defined by the ellipse

$$\left\{\alpha:(\alpha-\widehat{\alpha})'[\mathrm{Cov}\,(\widehat{\alpha})]^{-1}\,(\alpha-\widehat{\alpha})\le z_{0.025}^{2}\right\} \tag{9}$$

where $\alpha = (\alpha_0, \alpha_1)'$, $z_{0.025}{}^2 = 3.842$. Equivalently, the 95% joint confidence region consists of α values that satisfy

$$\frac{(\alpha_0-\widehat{\alpha}_0)^2\mathrm{var}\,(\widehat{\alpha}_1)-2\,(\alpha_0-\widehat{\alpha}_0)\,(\alpha_1-\widehat{\alpha}_1)\,\mathrm{cov}\,(\widehat{\alpha}_0,\widehat{\alpha}_1)+(\alpha_1-\widehat{\alpha}_1)^2\mathrm{var}\,(\widehat{\alpha}_0)}{\mathrm{var}\,(\widehat{\alpha}_0)\,\mathrm{var}\,(\widehat{\alpha}_1)-[\mathrm{cov}\,(\widehat{\alpha}_0,\widehat{\alpha}_1)]^2}\le z_{0.025}^{2} \tag{10}$$

Estimated variances for outcomes can be determined from Table 3, ie, $\widehat{\mathrm{var}}\,(\widehat{\alpha}_0)=[(\mathrm{UCB}-\widehat{\alpha}_0)\,/1.96]^2$ or $\widehat{\mathrm{var}}\,(\widehat{\alpha}_0)=[(\widehat{\alpha}_0-\mathrm{LCB})\,/1.96]^2$, and similarly for α_1. Using national evaluation estimates for past 30-day alcohol use (i.e., $\widehat{\alpha}_0=0.0261$, $\widehat{\alpha}_1=0.0219$, $\mathrm{var}\,(\widehat{\alpha}_0)=0.0000246$, $\mathrm{cov}\,(\widehat{\alpha}_0,\widehat{\alpha}_1)=0.0000128$, and $\mathrm{var}\,(\widehat{\alpha}1)=0.0000186$ gives the ellipse plotted in Figure 1 along with contours of power to show how power varies over the joint confidence region for (α_0, α_1). First note the ellipse represents the boundary of the joint 95% confidence region for $(\widehat{\alpha}_0,\widehat{\alpha}_1)$. Second, the box is the intersection of the individual 95% confidence intervals, ignoring the covariance between the 2 correlation estimates. Third, the numbers on the plot represent power to reject $H_0 : \delta_1 = 0$ in favor of $H_1 : \delta_1 \neq 0$, for the assumed values of probabilities described above. Fourth, the 3 bands within the ellipse indicate constant levels of power (77%, 80%, 83%) for different combinations of α_0 and α_1. Within the ellipse, power attains it highest value (85%) for the boundary values of ($\alpha_0 = 0.0210$, $\alpha_1 = 0.0250$), and its lowest value (76%) for the boundary values of ($\alpha_0 = 0.0311$, $\alpha_1 = 0.0187$). Thus, for a sample design with n = 48 and m = 15, the 95% joint confidence region for the 2 intracluster correlations indicates a range of power from 76% to 85%. This sensitivity analysis for power is a considerable refinement over a naive approach, represented by the box in Figure 1, that gives a range of power that is artificially low (71%) or high (90%). Given knowledge of the covariance of the 2 intracluster correlations, the joint elliptical confidence region approach represents a clear improvement in cluster trial planning.

Providing further rationale for use of the proposed 2 intracluster correlation sample size formula is the strong possibility that use of the posttest-only sample size formula when planning a pretest-posttest nested cross-sectional trial will underestimate the sample size needed to obtain a desired level of statistical power in a GEE time × condition analysis. Consider the ratio (r) of variances under the respective designs, var(d) based upon Equation 7 divided by var(d) based upon Equation 4. A ratio near 1 indicates that the formula for the posttest-only design is a good substitute for the pretest-posttest design formula. Figure 2 shows that generally $r > 1$ indicating that use of the posttest-only formula underestimates the required sample size. Using results for past 30-day alcohol use and $m = 15$, r ranges from 1.37 (corresponding to the highest value of power in the ellipse of Fig. 1) to 1.47 (based upon GEE estimates of intracluster correlations) to 1.56 (corresponding to lowest power). Figure 2 indicates that, in the case of ($\alpha_0 = 0.0210$, $\alpha_1 = 0.0250$), $r < 1$ only when $m \geq 30$; note $m = 30$ and $\alpha_0 = 0.0210$ gives $\phi = 1.61$, a relatively large design effect. Comparatively, the observed design effect for past 30-day use, calculated as $\phi = 1 + \widehat{\alpha}_0\left(\bar{m}-1\right)$, where $\bar{m}$ is the average number of youth from a community sampled at a timepoint, was 1.46. The fact that past 30-day use had the largest design effect among the 9 measures in the national evaluation suggests that observing a sufficiently large design effect in any cluster trial such that $r \leq 1$ appears unlikely.

DISCUSSION

The utility of any sample size formulae for cluster unit trials depends upon the availability of intracluster correlation estimates for various outcomes. Given their typically high cost, many cluster trials enroll a small number of clusters. However, in many government and foundationsponsored programs, such as the national evaluation, the resources available for the evaluation of the program may be distinct from funds for implementing the intervention, so a moderately large number of clusters may be studied. These larger studies offer a unique opportunity to report on the magnitude and precision of intracluster correlation estimates for health behavior outcomes.

Relative to equivalent general power methods for GEE,[21] the sample size formulae for a time × condition analysis of binary outcomes presented in this paper are easy to apply. This article emphasized sample size formulae based upon a planned GEE analysis using a logit link, whereas Preisser et al[12] emphasized similar formulae based upon an identity link. As a rule, the choice of sample size formula should be based upon the planned statistical analysis. However, because the 4-parameter interaction model in Equation 6 places no structure on the 4 time × condition probabilities, correlation estimates obtained with the logit link model (eg, those in Table 3) may be used in sample size formula in equation 8 for a planned logit analysis or in Formulae 3 and 9 of Preisser et al[12] for an analysis using the identity link. The difference in the respective formulae pertains to a difference in the definition of the effect of interest as a contrast of logits, or a contrast of proportions. Applying the formula based upon the identity link would have given $n = 50$ communities per condition in the previous section instead of $n = 48$; it is not always the case that larger sample sizes will be required for the identity link.

One practical obstacle in applying the proposed sample size methods is that information regarding variances of intracluster correlations (usually, in the form of confidence intervals or standard errors) are only occasionally published, and it seems less realistic that covariances of the 2 types of intracluster correlation will be available. If the 2 intracluster correlations are approximately equal, the problem may be circumvented by conducting a sensitivity analysis of power using the pretest-posttest formula under the assumption that $\alpha \equiv \alpha_0 = \alpha_1$; in this case, knowledge of a single intracluster correlation and its standard error are sufficient.

This article addressed the question of whether Formula 5 of the posttest-only design may be substituted for Formula 8 in planning a nonrandomized pretest-posttest nested cross-sectional cluster trial. Direct comparison of variances under the 2 designs showed that the posttest-only formula is generally not an appropriate substitute for the prettest-posttest design formula and will likely underestimate the required sample size and lead to an underpowered cluster trial. In other words, for nonrandomized cluster trials, pretest adjustment with a time × condition analysis is undertaken to address potential bias, at the price of loss of power. This is in contrast to the design and analysis of randomized cluster trials, where adjustment for pretest using a time × condition analysis may not be preferred because of the increased variance associated with Equation 7 relative to Equation 4. Rather, because communities tend to be similar across conditions due to randomization, covariate adjustment of cluster baseline response means may be undertaken to increase power.[22]

The Enforcing Underage Drinking Laws Program that provided the intracluster correlation estimates in Table 3 had limitations. The sample under-represented 19- and 20-year-old subjects compared with 16-, 17-, and 18-year-old subjects. Such selection bias may lead to biased intracluster correlation estimates. However, for the national evaluation, stratified analyses of intracluster correlation estimates (not reported) were similar across age groups. Another limitation is that the random digit dialing methodology used to conduct the survey is known to underrepresent ethnic/racial minorities and lower socioeconomic status individuals.[23] These limitations should be considered when deciding whether to use the reported intracluster correlations in planning a future cluster trial.

Finally, there are limitations with respect to the extended GEE and proposed sample size methods. The simple sample size formulae applied to the national evaluation data are applicable only to cluster trial study designs without matching or stratification. For these more complicated designs, use of the general GEE power method[21] is recommended. Finally, using the extended GEE to produce confidence intervals for intracluster correlations requires a large number of clusters (eg, 80 or more); otherwise, confidence intervals may suffer from severe under-coverage.[24] Small-sample bias adjustments to estimation of intracluster correlations,[25] and to their estimated variances and covariances by extension of corrections for ordinary GEE[26,27] may broaden the applicability of these methods. However, because the validity of estimating equation methods depends upon the assumption of asymptotic normality of parameter estimates, construction of (possibly asymmetric) confidence intervals based upon resampling strategies may be a better choice for small samples. In articular, bootstrap methods,[28] though more computationally intensive than the GEE approach, are often easy to implement.[29]

ACKNOWLEDGMENTS

We are grateful to 2 anonymous referees for providing helpful comments.

The research was supported by grants from the Office of Juvenile Justice and Delinquency Prevention, U. S. Department of Justice 98-AH-F8-0101 (S-1, S-2, S-3, S-4), and 2005-AH-FX-K011.

APPENDIX

Equivalency of Sample Size Formula 5 for the Posttest-Only Analysis to the General GEE Sample Size Method

Rochon[21] presents a general power analysis method for GEE regression coefficients for arbitrary working correlation and link and variance functions. Now, Equation 5 is the sample size formula for the summary statistic approach to power for the test statistic d/var(d), which in sufficiently large samples is normally distributed and equivalent to

$$Z=\frac{\text{logit}\,(\pi_{11})-\text{logit}\,(\pi_{21})}{\sqrt{\frac{\phi}{nm}\left(\frac{1}{v_{11}}+\frac{1}{v_{21}}\right)}}$$

The following arguments show that the GEE Wald χ^2 statistic, Q_W in Expression 5 of Rochon,[21] is equal to Z^2 thus demonstrating equivalency of the 2 power analysis approaches. This first step deduces the GEE estimator $\widehat{\beta}$ in Equation 2 of Rochon[21] for the simple logistic model in Equation 3, binomial variance function ν_{h1}, and "exchangeable" working correlation matrix $R = \phi(J/m) + (1 - \alpha_0) \times (I - J/m)$, where I is the $m \times m$ identity matrix and J is the $m \times m$ matrix of 1's. Define **0** and **1** as $m \times 1$ vectors of 0's and 1's, respectively. Following Equation 2 of Rochon,[21] the GEE estimator for $\beta = (\beta_0, \delta_0)$is

$$\widehat{\beta}=\left[\sum_{h=1}^{2}X_h'W_hX_h\right]^{-1}\left[\sum_{h=1}^{2}X_h'W_h\ (1\mathrm{g_h})\right]$$

where $X_1 = [\mathbf{1}, \mathbf{1}]$ and $\chi_2 = [\mathbf{1}, \mathbf{0}]$, $W_\mathrm{h} = \Delta_\mathrm{h}^{-1}\ \Delta_\mathrm{h}$, $\Delta_\mathrm{h} = \text{diag}\{\sqrt{v_{h1}}\}$, and $g_\mathrm{h} = \text{logit}(\pi_{\mathrm{h}1})$ for h = 1, 2. Matrix computations and the result $R^{-1} = J/(m\phi) + (I - J/m)/(1 - \alpha_0)$ lead to $\widehat{\beta} = (g_2, g_1 - g_2)'$ and, from Equation 3 of Rochon,[21] its model-based variance estimator is

$$\text{COV}_{MB}\left(\widehat{\beta}\right)=\frac{\phi}{nm}\begin{bmatrix}1/v_{21} & -1/v_{21}\\ -1/v_{21} & 1/v_{21}+1/v_{11}\end{bmatrix}.$$

Finally, the hypothesis $H_0 : \delta_0 = 0$ can be expressed as $H_0 : \boldsymbol{H}\beta = 0$, where $\boldsymbol{H} = [0, 1]$. The Wald chi-square test statistic is

$$Q_W = \left(H\widehat{\beta}\right)' \left[H \mathrm{cov}_{MB} H'\right]^{-1} \left(H\widehat{\beta}\right) = Z^2$$

proving the equivalency.

REFERENCES

1. Donner, A.; Klar, N. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold; London: 2000.
2. Murray, DM. Design and Analysis of Group-Randomized Trials. Oxford University Press; New York: 1998.
3. Klar N, Donner A. Current and future challenges in the design and analysis of cluster randomization trials. Stat Med 2001;20:3729–3740. [PubMed: 11782029]
4. Murray DM, Short B. Intraclass correlation among measures related to alcohol use by young adults: estimates, correlates and applications in intervention studies. J Studies Alcohol 1995;56:681–694.
5. Murray DM, Short B. Intraclass correlation among measures related to alcohol use by school aged adolescents: estimates, correlates and applications in intervention studies. J Drug Education 1996;26:207–230.
6. Siddiqui O, Hedeker D, Flay BR, Hu FB. Intraclass correlation estimates in a school-based prevention study. Am J Epidemiol 1996;144:425–433. [PubMed: 8712201]
7. Klar N, Darlington G. Methods for modelling change in cluster randomization trials. Stat Med 2004;23:2341–2357. [PubMed: 15273952]
8. Murray DM, Hannan PJ, Wolfinger RD, Baker WL, Dwyer JH. Analysis of data from group-randomized trials with repeat observations on the same groups. Stat Med 1998;17:1581–1600. [PubMed: 9699231]
9. Bellamy SL, Gibberd R, Hancock L, Howley P, Kennedy B, Klar N, Lipsitz S, Ryan L. Analysis of dichotomous outcome data for community intervention studies. Stat Methods Med Res 2000;9:135–159. [PubMed: 10946431]
10. Sashegyi AI, Brown KS, Farrell 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;152:1192–1200. [PubMed: 11130626]
11. Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988;44:1033–1048. [PubMed: 3233244]
12. 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;22:1235–1254. [PubMed: 12687653]
13. Ukoumunne OC, Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. Stat Med 2001;20:417–433. [PubMed: 11180311]
14. Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials. Stat Med 1994;13:61–78. [PubMed: 9061841]
15. Wolfson, M.; Altman, D.; DuRant, R., et al. National Evaluation of the Enforcing Underage Drinking Laws Program: Year 4 Report. Wake Forest University School of Medicine; Winston-Salem, NC: 2004 [February 2, 2007]. Available at: http://www.phsintranet.wfubmc.edu/EUDL2/pubs.cfm.
16. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
17. Murray DM, Clark MH, Wagenaar AC. Intraclass correlations from a community-based alcohol prevention study: the effect of repeat observations on the same communities. J Studies Alcohol 2000;61:881–890.
18. Murray DM, Lee Van Horn M, Hawkins JD, Arthur MW. Analysis strategies for a community trial to reduce adolescent ATOD use: a comparison of random coefficient and ANOVA/ANCOVA models. Contemporary Clin Trials 2006;27:188–206.
19. Slymen DJ, Elder JP, Litronik AJ, Ayala GX, Campbell NR. Some methodologic issues in analyzing data from a randomized adolescent tobacco and alcohol use prevention trial. J Clin Epidemiol 2003;56:332–340. [PubMed: 12767410]

20. Zink, R.; Preisser, JS.; Lu, B.; Perin, J. GEECORR: A SAS macro implementing estimating equations for binary data. [May 15, 2007]. Available at: http://www.bios.unc.edu/~jpreisse/personal/software.htm.

21. Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. Stat Med 1998;17:1643–1658. [PubMed: 9699236]

22. Stevens J, Murray DM, Catellier DJ, Hannan PJ, Lytele LA, Elder JP, Young DR, Simons-Morton DG, Webber LS. Design of the Trial of Activity in Adolescent Girls (TAAG). Contemporary Clinical Trials 2005;26:223–233. [PubMed: 15837442]

23. Blumberg SJ, Luke JV, Cynamon ML. Telephone coverage and health survey estimates: evaluating the need for concern about wireless substitution. Am J Public Health 2006;96:926–931. [PubMed: 16571707]

24. Evans BA, Feng Z, Peterson AV. A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. Stat Med 2001;20:3353–3373. [PubMed: 11746323]

25. Sharples K, Breslow N. Regression analysis of correlated binary data: some small samples results for the estimating equation approach. J Statl Computation Simulation 1992;42:1–20.

26. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics 2001;57:126–134. [PubMed: 11252587]

27. Lu B, Preisser JS, Qaqish BF, et al. A comparison of two bias-corrected covariance estimators for generalized estimating equations. Biometrics. in press. Published online 2 Mar 2007; doi:10.1111/j.1541−0420.2007.00764.x

28. Sherman M, le Cessie S. A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. Comm Stat: Simulation 1997;26:901–925.

29. Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics 2002;58:1020–1027. [PubMed: 12495158]

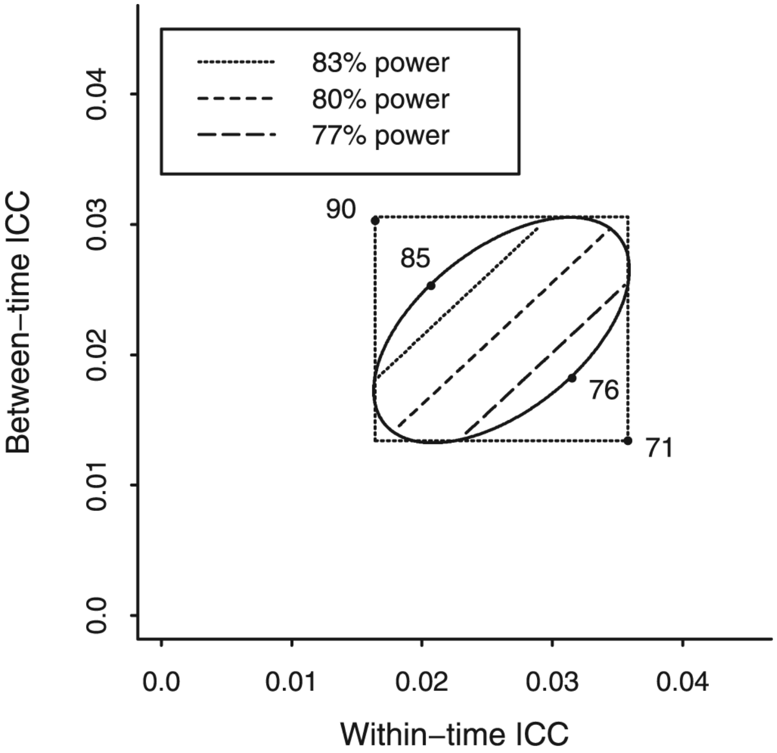

FIGURE 1.

Power for detecting a difference in prevalence of past 30-day alcohol use of 0.10 between control and intervention conditions with a pretest–posttest nested cross-sectional design with n = 48 and m = 15. The diagonal lines indicate pairs of within-time and between-time intracluster correlation values that give a fixed level of power and that are useful values of correlations for planning a future trial as they lie in the ellipse that defines a 95% joint confidence region for the intracluster correlations.

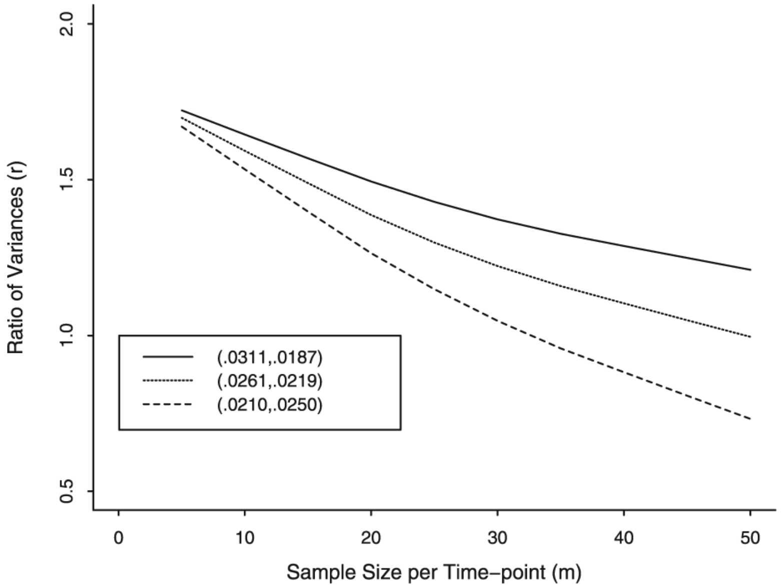

FIGURE 2.
Ratio of variances of intervention effect from pretest–posttest to posttest-only design by sample size for 3 values for pairs of within-time and between-time intracluster correlations.

TABLE 1

Description of Alcohol Use Measures From the Youth Survey of the Enforcing Underage Drinking Laws Program, 1998−2004

Measure	Description
Binge drinking	“Think back over the last 2 weeks. How many times have you had 5 or more drinks in a row?” Respondents who answered zero were coded “0” whereas those who answered one or more occasions were coded “1.”
DWI driving	Among ever-drinkers who reported ever driving, “During the last 30 days, how many times (if any) have you driven after drinking 2 or more drinks in an hour or less?” Those reporting zero, never drinkers, and ever drinkers who had never driven were coded “0”; those replying one or more occasions were coded “1.”
Past 30-day alcohol use	“When was the last time you drank alcohol?” respondents who answered “in the last 30 days” were coded “1.” Respondents who answered that they had consumed alcohol but not in the last 30 days, as well as never drinkers, were coded “0.”
Past 7-day use	Similar to “*Past 30-day alcohol use*”; response “in the last 7 days” coded “1.”
Attempt to purchase alcohol	“In the last 30 days, how many times did you try to buy alcohol from a bar, restaurant, or store (whether you were successful or not)?” Respondents who answered zero were coded “0”; those who answered one or more occasions were coded “1.”
Nonviolent consequences due to alcohol use	Experienced any of the following after they had been drinking: being cited or arrested for drinking, possessing alcohol, trying to buy alcohol, being cited or arrested for driving under the influence of alcohol; missing any school due to drinking; being warned by a friend about your drinking; passing out; being unable to remember what happened while drinking; breaking or damaging something; having a headache or hangover; being punished by your parents or guardian; and having sex without using birth control. Respondents who reported experiencing any these in the past year were coded “1”; others were coded “0.”
Perception of alcohol use among peers	“How many of your friends do you think have had any alcohol to drink in the last 30 days? Would you say none, a few, some, most, or all”? Respondents who reported that most of all or their friends had consumed alcohol in the past 30 days were coded “1”, while all others were coded “0.”
Perception of getting caught by police	“If you had been drinking, how likely would it be for the police to catch you? Very likely, somewhat likely, not very likely, or not at all likely?”, coded “1” if respondent answered “very likely” or “somewhat likely”; else coded “0.”
Commercial source of alcohol	“The last time you drank any alcohol, how did you get the alcohol?” those who answered that they obtained alcohol from a “commercial source” (businesses such as alcohol outlets, restaurants, and bars) were coded as “1.” Those who answered that they were given alcohol by their friends, family members, coworkers, acquaintances, or strangers at home or at events were coded “0.”

TABLE 2

Sex, Race, and Age of Participants in the Youth Survey of the Enforcing Underage Drinking Laws Program for Control and Intervention Communities (All Rounds), 1998−2004

	Year 1		Year 2		Year 3	
	Control (n = 1835)*	Intervention (n = 1784)*	Control (n = 1825)*	Intervention (n = 1822)*	Control (n = 1825)*	Intervention (n = 1775)*
Male; %	49.1	50.0	50.3	51.2	48.4	51.3
Race; %						
White	85.4	87.5	86.6	88.1	83.1	87.6
Black	6.6	5.2	6.4	5.4	8.5	6.0
Other	8.0	7.3	7.0	6.4	8.7	6.5
Age (yrs); %						
16	32.0	30.8	31.1	30.4	30.2	28.2
17	31.1	32.0	31.8	31.2	29.4	29.8
18	17.6	18.0	17.1	19.7	19.2	21.2
19	11.6	10.9	11.3	10.4	12.1	12.9
20	7.7	8.4	8.7	8.3	9.0	8.0
Community sample; mean†	18.0	17.8	17.9	18.2	17.7	18.0

* Total number of youth sampled across communities.

† Mean number of youth sampled per community.

TABLE 3

Within-Time (α_0) and Between-Time (α_1) Intracluster Correlation Estimates (ICC) of Youth Alcohol Use Measures and Their 95% Confidence Intervals (CI) From the Youth Survey of the Enforcing Underage Drinking Laws Program, 1998–2004

Measure (Prevalence[*]) and Independent Variables[†]	**Within-Time ICC (95% CI)**	**Between-Time ICC (95% CI)**	**Covariance[‡]**
Binge drinking (0.1656)			
Design variables	0.0185 (0.0048 to 0.0322)	0.0169 (0.0088 to 0.0251)	0.0000200
Age, sex	0.0193 (0.0060 to 0.0325)	0.0172 (0.0081 to 0.0262)	0.0000229
Community variables	0.0192 (0.0053 to 0.0330)	0.0166 (0.0069 to 0.0262)	0.0000226
DWI drive (0.0499)			
Design variables	0.0027 (−0.0037 to 0.0092)	0.0074 (0.0019 to 0.0129)	0.00000252
Age, sex	0.0038 (−0.0032 to 0.0109)	0.0070 (0.0010 to 0.0131)	0.00000438
Community variables	0.0038 (−0.0033 to 0.0109)	0.0071 (0.0010 to 0.0133)	0.00000452
Past 30-day alcohol use (0.3939)			
Design variables	0.0261 (0.0164 to 0.0358)	0.0219 (0.0134 to 0.0303)	0.0000128
Age, sex	0.0251 (0.0156 to 0.0346)	0.0217 (0.0133 to 0.0301)	0.0000117
Community variables	0.0234 (0.0139 to 0.0328)	0.0197 (0.0112 to 0.0282)	0.0000120
Past 7-day alcohol use (0.2174)			
Design variables	0.0190 (0.0092 to 0.0288)	0.0158 (0.0088 to 0.0228)	0.00000745
Age, sex	0.0204 (0.0106 to 0.0301)	0.0156 (0.0082 to 0.0229)	0.00000785
Community variables	0.0197 (0.0099 to 0.0295)	0.0143 (0.0064 to 0.0222)	0.00000880
Attempt to purchase alcohol (0.0677)			
Design variables	0.0191 (0.0096 to 0.0286)	0.0194 (0.0093 to 0.0295)	0.0000128
Age, sex	0.0178 (0.0069 to 0.0286)	0.0184 (0.0086 to 0.0281)	0.0000164
Community variables	0.0088 (−0.0003 to 0.0180)	0.0098 (0.0030 to 0.0165)	0.00000582
Nonviolent consequences to alcohol use (0.3609)			
Design variables	0.0112 (0.0039 to 0.0184)	0.0115 (0.0056 to 0.0173)	0.00000500
Age, sex	0.0106 (0.0054 to 0.0169)	0.0111 (0.0054 to 0.0169)	0.00000451
Community variables	0.0087 (0.0022 to 0.0152)	0.0092 (0.0040 to 0.0145)	0.00000309
Perception of alcohol use among peers (0.5967)			
Design variables	0.0222 (0.0124 to 0.0319)	0.0200 (0.0116 to 0.0284)	0.0000131
Age, sex	0.0208 (0.0115 to 0.0300)	0.0196 (0.0115 to 0.0276)	0.0000116
Community variables	0.0204 (0.0112 to 0.0296)	0.0192 (0.0110 to 0.0273)	0.0000117
Perception of getting caught by police (0.3988)			
Design variables	0.0114 (0.0033 to 0.0195)	0.0125 (0.0061 to 0.0190)	0.00000672
Age, sex	0.0118 (0.0035 to 0.0200)	0.0127 (0.0062 to 0.0192)	0.00000694
Community variables	0.0087 (0.0010 to 0.0164)	0.0097 (0.0038 to 0.0157)	0.00000492
Commercial source of alcohol (0.0730)			
Design variables	0.0163 (0.0036 to 0.0290)	0.0157 (0.0049 to 0.0265)	0.0000206
Age, sex	0.0164 (0.0040 to 0.0288)	0.0159 (0.0058 to 0.0262)	0.0000190
Community variables	0.0069 (−0.0026 to 0.0165)	0.0064 (−0.0004 to 0.0132)	0.0000339

[*] Observed overall prevalence.

[†] Design variables are condition, time, round and their pairwise interactions. Community variables are percent of household with female head with no husband, percent foreign born, and median income.

[‡] Covariance between $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$.