Austin Biometrics and Biostatistics

Open Access

Austin Publishing Group

Special Article - Biostatistics Theory and Methods

# Simulating Clustered and Dependent Binary Variables

**Aobo Wang and Roy T Sabo***
Department of Biostatistics, Virginia Commonwealth University, USA

***Corresponding author:** Sabo RT, Department of Biostatistics, Virginia Commonwealth University, 830 East Main Street, Richmond, VA 23298-0032, USA

**Received:** June 01, 2015; **Accepted:** June 11, 2015; **Published:** June 19, 2015

**Abstract**

Dependent binary data can be simply simulated using the multivariate normal- and multinomial sampling-based approaches. We extend these methods to simulate dependent binary data with clustered random effect structures. Several distributions are considered for constructing random effects among cluster-specific parameters and effect sizes, including the normal, uniform and beta distributions. We present results from simulation studies to show proof of concept for these two methods in creating data sets of repeated-measure binary outcomes with clustered random effect structures in various scenarios. The simulation studies show that multivariate normal- and multinomial sampling approaches can be successfully adapted to simulate dependent binary data with desired random effect structures.

**Keywords:** Dependent binary data; Clustered random effect; Simulated data

## Abbreviations

MVN: Multivariate Normal; CDF: Cumulative Distribution Function; PDF: Probability Density Function; MS: Multinomial Sampling

## Introduction

Methods for simulating dependent binary outcomes are often required for the assessment of statistical methodologies suitable for repeated measure study designs with dichotomous outcomes. Such simulation techniques can also be useful in determining required sample sizes for longitudinal study designs featuring binary measurements. Emrich and Piedmonte [1] developed a gold-standard method for simulating dependent binary outcomes based on the multivariate normal distribution. Kang and Jung [2] introduced an approach based on the multinominal distribution of all possible combinations of the binary outcomes. Both of these approaches were extended to account for modeling dependencies with odds ratios in Sabo et al. [3].

While useful for repeated-measures or multiple-outcome studies, these methods require expansion if they are to be used in more complicated situations. For instance, certain research studies feature inherent clustering, where groups of subjects exist in natural clusters or groups. Examples include studies of school-age children attending various class rooms or schools [4], or primary care patients who attend one of several primary care facilities [5], the latter of which also features patients nested within primary care physicians, who are in turn nested within primary care practices that are nested within larger health care systems. The previously mentioned simulation approaches cannot incorporate this type of complexity without amendment and are unsuitable as currently constructed to simulate clustered repeated measure data that would mimic such a scenario.

In this manuscript, we extend the multivariate normal- and multinomial sampling-based approaches for simulating dependent binary outcomes to also incorporate a desired cluster structure. This extension requires probabilistically generating parametric simulation templates for each of the desired cluster levels or combinations. Several simple probability distributions are used to exemplify the process of establishing the cluster-specific parameters and effect sizes, including the normal, uniform and beta distributions. The rest of this manuscript is outlined as follows. The two simulation methods are briefly described in the next Section, and are extended to account for a desired cluster structure. The performances of these extensions are then examined through simulation studies. A brief discussion concludes the manuscript.

## Materials and Methods

### Simulation methodologies: Multivariate normal approach

The simulation approach by Emrich and Piedmonte [1] utilizes the multivariate normal distribution to generate vectors exhibiting desired dependence levels, which are then categorized into binary observations. The process begins by using the desired pairwise correlations $\rho_{ij}$ between binary measures $Y_i$ and $Y_j$ with marginal probabilities $p_i=P(Y_i=1)$ and $p_j=P(Y_j=1)$ to solve for a bivariate correlation $r_{ij}$ using the bivariate normal Cumulative Distribution Function (CDF)

$$\Phi[z(p_i),z(p_j),r_{ij}]=\rho_{ij}(p_iq_ip_jq_j)^{1/2}+p_ip_j, \quad (1)$$

where $z(p)$ is the $p^{th}$ percentile of the standard normal distribution and $q = 1 - p$. Odds ratios could be used in place of correlations by replacing the right-hand side of Equation(1) with the Plackett copula [6] $C(p_i,p_j,\Psi_{ij})$, where $\Psi_{ij}$ is the desired odds ratio, as shown in Sabo *et al.*[3]. The values $r_{ij} \forall i \neq j$ are then placed into a correlation matrix $R$ and used to simulate a $k \times 1$ multivariate normal vector $z= (z_1,\ldots,z_k)^T \sim MVN(0,R)$. Binary observations are then created by classifying each element of $z$ by letting $Y_i=1$ if $z_i \leq z(p_i)$ and $Y_i=0$ otherwise. This process can be repeated by generating and classifying $n$ such vectors to create the desired simulated sample.

### Simulation methodologies: Multinomial approach

The multinomial-based simulation method introduced by Kang and Jung [2] uses a multinomial distribution of all possible combinations of dependent binary outcomes, which can be created




**Citation:** Wang A and Sabo RT. Simulating Clustered and Dependent Binary Variables. Austin Biom and Biostat. 2015;2(2): 1020.

through the joint and marginal probabilities, along with the desired correlation. Given a desired correlation $\rho_{ij}$ between binary variables $Y_i$ and $Y_j$ with desired marginal probabilities $p_i$ and $p_j$, we first calculate the joint probability $p_{ij}$ using the following expression.

$$p_{ij} = p_i p_j + \rho_{ij}\sqrt{p_i q_i}\sqrt{p_j q_j} \quad (2)$$

Note that if odds ratios are used instead of correlations, then $p_{ij}$ can be solved for by inserting the desired odds ratio $\Psi_{ij}$ and marginal probabilities $p_i$ and $p_j$ into the Plackett copula, as described in Sabo *et al.* [3]. Note that whether correlations or odds ratios are used to model dependence, the remainder of the multinomial-based approach is identical after the pair-wise joint probabilities $p_{ij}$ are calculated.

If three or more dependent binary measures are to be simulated, then higher order joint probabilities must be calculated. Let $p_{ijk}$ represent the joint probability $P(Y_i=1,Y_j=1,Y_j=1)$, which is not uniquely defined by the marginal probabilities and the correlation. As shown in Chaganty and Joe [7], the minimum and maximum $p_{ijk}$ are defined as follows,

$$p_{ijk,L}=max\{0,p_{ij}+p_{ik}-p_i,p_{ij}+p_{jk}-p_j,p_{ik}+p_{jk}-p_k\} \quad (3)$$

$$p_{ijk,U}=max\{p_i,p_j,p_k,1-p_i-p_j-p_k+p_{ij}+p_{ik}+p_{jk}\}$$

where any value $p_{ijk}\in[p_{ijk,L},p_{ijk,U}]$ leads to a valid probability density function with the desired marginal probabilities and dependence level. Though any value in this range is appropriate, we take the midpoint $p_{ijk}=(p_{ijk,L}+p_{ijk,U})/2$. Higher order joint probabilities in cases of four or more dependent binary observations can be determined in a similar manner, though the calculations become more tedious as the number of observations increases.

These quantities are used to calculate the multinomial Probability Density Function (PDF) of all combinations of outcomes, which for the two-variable case are shown in the first two columns of Table 1. The CDF is created by progressively summing the values of the PDF, where the subscripts on $P$ indicate whether each binary outcome is successful, with 1 for success and 0 for failure. For example, $P_{01}=P(Y_1=0,Y_2=1)$. After the CDF is determined, a random number $u\sim Uniform\ [0,1]$ is simulated, and the simulated observations are generated based on the decision rules based on the CDF, as shown in the last two columns of Table 1. For example, if $P_{11}<u<P_{11}+P_{10}$, then the observation is recorded as $Y_1=1$ and $Y_2=0$, or simply as 10. This process can be repeated to generate a sample of $n$ dependent binary outcomes. A similar approach – outlined in Kang and Jung [2] and Haynes et al. [8] – can be used in cases of three or more dependent binary outcomes.

## Accounting for random effects by generating the simulation templates

For the two simulation approaches discussed in Section 2, we simulate a set of binary data representative of a single population by repeating either process $n$ times using a single simulation template, which consists of all desired marginal probabilities $p_i, i=1,\ldots,k$ and pair wise correlations $\rho_{ij}$ (or odds ratios $\Psi_{ij}$). To generate two or more groups of simulated binary observations, where groups are differentiated by either different marginal probabilities, dependencies, or both, the simulation approach is repeated separately for each group with the desired simulation template.

**Table 1:** Two-variable PDF, CDF and decision rules for multinomial approach.

| PDF | CDF | Decision Rule and Simulated Outcome | |
|---|---|---|---|
| $P_{11}=p_{12}$ | $P_{11}$ | $U\le P_{11}$ | 11 |
| $P_{10}=p_1-p_{12}$ | $P_{11}+P_{10}$ | $P_{11}<U\le P_{11}+P_{10}$ | 10 |
| $P_{01}=p_2-p_{12}$ | $P_{11}+P_{10}+P_{01}$ | $P_{11}+P_{10}<U\le P_{11}+P_{10}+P_{01}$ | 01 |
| $P_{00}=1-p_1-p_2+p_{12}$ | $P_{11}+P_{10}+P_{01}+P_{00}$ | $U>P_{11}+P_{10}+P_{01}$ | 00 |

Expanding the multivariate- and multinomial-based approaches to account for random effects (say from clustering) requires only a simple extension of the process used when simulating binary observations for multiple groups. As a motivating example, let's assume we want to simulate $M$ clusters of $n$ samples of two correlated binary measures. Let's further assume that those two measures have marginal probabilities that vary across the $M$ clusters in such a way that the averages are $p_1=\pi_1$ and $p_2=\pi_2$ and the corresponding cluster variances for those rates are $\sigma_1$ and $\sigma_2$.

First we assume that the marginal probability for each binary measure has some probability distribution $p_i\sim f(\theta_i)$, where $f(o)$ is some probability mass or density function and $\theta$ is some parameter (possibly vector-valued) selected such that $E(p_i)=\int p_i f(\theta_i)dy=\pi_i$ and $V(p_i)=\int(p_i-E(p_i))^2 f(\theta_i)dy=\sigma_i$ for group $i=1,2$. If we desire $M$ clusters of simulated observations, then we simulate $M$ marginal probabilities $p_{1,m}$ and $p_{2,m}$ from $f(\theta_1)$ and $f(\theta 2)$, respectively, for $m=1,\ldots,M$. For cluster $m$, we simulate the desired number of dependent binary observations using $p_{1,m}$ and $p_{2,m}$ and the desired dependence level $\rho_{12}(or\psi_{12})$. This process is repeated for $m=1,\ldots,M$, and the resulting $M$ clusters of simulated data will *on average* exhibit a distribution of marginal probabilities centered around $\pi_1$ and $\pi_2$, though the cluster-specific marginal means will vary according to $\sigma_1$ and $\sigma_2$, thus achieving the desired level of clustering.

In the previous scenario, the marginal probabilities were given probability distributions and themselves simulated $M$ times to achieve a clustering effect. An equivalent approach would be to simulate $p_1\sim f(\theta_1)$ probabilistically to achieve a desired mean and variance for the first marginal mean across clusters, and then simulate some $\delta\sim g(\gamma)$ and define $p_2=p_1+\delta$, where $g(o)$ is some probability distribution not necessarily of the same form as $f(o)$, and $\gamma$ is some parameter (possibly vector-valued) such that $E(\delta)=\int\delta g(\gamma)d\delta$ yields the desired difference between $p_1$ and $p_2$ with some desired cluster variability $V(\delta)=\int(\delta-E(\delta))^2 g(\gamma)d\delta$.

This approach extends naturally to more complicated scenarios, including cases of two or more clustering factors, or even nested factors. The unifying theme is that data are simulated uniquely for each combination of clusters, mainly through parametric templates that are probabilistically generated for each combination. For example, in the case of hierarchical clustering, where one factor is nested within the levels of another, the parameters $\theta_i$ used to simulate the parameter values, which are used to simulate data for each cluster can be probabilistically determined. Further, the researcher has much discretion in selecting how those factors or levels affect the particular probability distribution and parameters used to simulate the simulation template for each cluster. The dependence levels between the binary outcomes can also be made to be cluster- or level-dependent, provided a distribution is selected that offers control in selecting the desired dependence while also ensuring the proper support.

### Distribution examples

We consider three examples of distributions that can be used in this simulation process, understanding that there are alternative and potentially more suitable options available. The only requirement is that the support of the distribution must either be equal to *[0,1]*, be a proper subset of *[0,1]*, or have a reasonably low probability of occurring outside *[0,1]*. A simple choice would be to simulate the marginal probabilities from a uniform distribution such that $p_{i,m} \sim Uniform[\theta_{i1},\theta_{i2}]$ for $m=1,\ldots,M$ clusters and $i=1,\ldots,k$ binary outcomes, where the midpoint of $\theta_{i1}$ and $\theta_{i2}$ yields the desired marginal mean $\pi_i$. In this case the inter-cluster variability in marginal probabilities can be controlled by increasing or decreasing the difference $\theta_{i2}$-$\theta_{i1}$, making it wider for greater variability and narrower for less variability. In this case the *Uniform* parameters can be selected such that $p_{i,m} \in [0,1] \forall i,m$.

Another example would be to simulate the marginal probabilities from a beta distribution such that $p_{i,m} \sim Beta[\alpha_i,\beta_i]$ for $m=1,\ldots,M$ clusters and $i=1,\ldots,k$ binary outcomes, where shape parameters $\alpha_i$ and $\beta_i$ are selected so that the mode is equal to the desired marginal probability (i.e. $(\alpha_i - 1)(\alpha_i + \beta_{i-2)} = \pi_i$). There are infinite pairings of the shape parameters that give the same mode, so the inter-cluster variability in the marginal probabilities is controlled by making both $\alpha_i$ and $\beta_i$ larger (for less variability) or smaller (for more variability). Since the support of the *Beta* distribution matches that of proportions and probabilities, we are assured that $p_{i,m} \in [0,1] \forall i,m$.

The final example we consider is to simulate marginal probabilities from a normal distribution with low variance such that $p_{i,m} \sim Normal\left(\pi_i, \sigma_i^2\right)$, where $\pi_i$ is the desired $i^{th}$ marginal probability and $\sigma_i^2$ is the desired inter-cluster variation, often these should be set to a common value $\sigma^2$. While this choice of distribution has infinite support, the variance $\sigma_i^2$ can be made small enough to all but ensure values are greater than 0 and less than 1 while simultaneously providing the desired variability. The simulations can also be truncated so that *p=0.001* if the simulated value is less than *0* and *p=0.999* if it is greater than *1*. Using the normal distribution also has the advantage of providing more direct control of the inter-cluster variability in marginal proportions as compared to the *Uniform* and *Beta* distributions.

### Extension to existing approaches

For the multivariate normal-based approach, the simulated marginal probabilities $p_{1,m},\ldots,p_{k,m}$ for clusters $m=1,\ldots,M$ are matched with the desired dependence levels $\rho_{ij}(or \psi_{ij})$ and are used in Equation 1 separately for each cluster. At this point, the process continues as stated in Section 2.1. Likewise, for the multinomial-based approach, the simulated marginal probabilities are matched with the desired dependence levels and used to determine the joint pair wise probabilities in Equation 2. Thereafter, the multinomial approach continues as stated in Section 2.2.

## Results and Discussion

### Simulation study

Here the performance of the multivariate normal and multinomial approaches to simulating dependent binary data with random effects is examined through simulation studies. The first case illustrates the simple situation where we simulate *k=2* dependent binary outcomes over *M=20* clusters. A second case looks at the situation where we simulate *k=2* dependent binary outcomes over *M=20* clusters, each consisting of both treatment and control subjects. For each case we assume the correlation between the two outcomes is $\rho_{12}=0.2$, irrespective of group and cluster. Sample size was fixed at *n=100* subjects per cluster. For both cases we also investigate the use of the, *Uniform*, *Beta* and *Normal* distributions for generating the simulation templates and incorporating the cluster-level variability. A total of *500* data sets was created for each combination of distribution and simulation method, and are used to estimate the average overall marginal probability for each measure, the standard error of those means, the average effect size (and standard error) for the case-specific hypothesis test, the empirical power for the case-specific hypothesis test, the mean and standard error of the inter-cluster variability, the mean estimated correlation, and the percentage of data sets for which the desired model converged. SAS (version 9.4, Cary, NC, USA) was used to simulate data and fit generalized Linear Mixed Models using the IML and GLIMMIX procedures, respectively.

### Case 1: Clustered, One-Group, Repeated-Measure Study

In this case we simulate *k=2* binary outcomes over *M=20* clusters, where the global marginal probabilities for the two outcomes are $p_1=0.25$ and $p_2=0.45$, indicating that the rate of our simulated outcome increases by *0.20* after some time (possibly after an intervention). To incorporate inter-cluster variance we simulate the marginal probabilities according to the specifications listed in Table 2. Here we see that: the midpoints of the two *Uniform* distributions are *(0.15+0.35)/2=0.25* and *(0.34+0.55)/2=0.45*, respectively; the modes of the two *Beta* distributions are *(11-1)/(11+31-2)=10/40=0.25* and *(10-1)/(10+12-2)=9/20=0.45*, respectively; and the means of the two normal distributions are *0.25* and *0.45*; in each case matching the target levels. These values also imply that the inter-cluster variability in the marginal means is *0.0033* for both measures with the *Uniform* distribution, *0.0045* and *0.0178* for the *Beta* distribution, and *0.01* for both measures with the *Normal* Distribution. The intended model is fit with a fixed two-level "time" effect, a cluster-level random effect to account for the inter-cluster variation, and a subject-level random effect to account for the correlation between the measures. The null

**Table 2:** Simulation template for case one.

| | Marginal | Distribution | | |
|---|---|---|---|---|
| Time | Mean | Uniform | Beta | Normal |
| 1 | $P_1$ | *U[0.15,0.35]* | *Beta(11,31)* | $N(0.25,0.1^2)$ |
| 2 | $P_2$ | *U[0.35,0.55]* | *Beta(10,12)* | $N(0.45,0.1^2)$ |

**Table 3:** Simulation results for case one.

| Dist. | Approach | $P_1$ (SE) | $P_2$ (SE) | $P_2$-$P_1$ (SE) | Cluster Random Effect (SE) | P (SE) | % Converged |
|---|---|---|---|---|---|---|---|
| *Uniform* | MS | 0.248 (0.015) | 0.449 (0.015) | 0.201 (0.023) | 0.034 (0.020) | 0.192 (0.021) | 99.0% |
| | MVN | 0.250 (0.016) | 0.449 (0.017) | 0.199 (0.022) | 0.033 (0.018) | 0.191 (0.023) | 100% |
| *Beta* | MS | 0.257 (0.017) | 0.453 (0.025) | 0.196 (0.030) | 0.075 (0.035) | 0.181 (0.023) | 99.8% |
| | MVN | 0.261 (0.019) | 0.457 (0.026) | 0.196 (0.031) | 0.076 (0.033) | 0.180 (0.022) | 100% |
| *Normal* | MS | 0.247 (0.026) | 0.447 (0.024) | 0.201 (0.034) | 0.102 (0.044) | 0.170 (0.025) | 100% |
| | MVN | 0.247 (0.025) | 0.450 (0.020) | 0.203 (0.035) | 0.100 (0.045) | 0.170 (0.023) | 100% |

hypothesis is no difference over time *(i.e.* $H_0$*:*$(p_1=p_2)$ against a two-sided alternative.

The aggregate results over the *500* simulations for case 1 are found in Table 3. Here we see that both the MS and MVN simulation approaches were accurate in reproducing the marginal proportions $p_1$ and $p_2$, as well as the difference $\delta=p_2-p_1$. We see that the MS and MVN approaches everywhere provided similar estimates and standard errors. The variability of these estimates is low and is also comparable between approaches. The cluster random effects averaged over all simulations are also provided; note these will not necessarily correspond to the theoretical inter-cluster variances stated earlier as these are model-derived and based on linked expectations in the generalized linear mixed model framework. The target correlation *(ρ=0.2)* was also achieved by both methods, with reasonably small variance. The MS approach produced data sets that converged at least 99.0%, while the MVN approach always converged. The empirical powers for testing the null hypothesis of no difference in change over time between the two groups for the both the multivariate normal and multinomial approaches were >99.9% for each of the three distributions (not shown in Table 3).

While comparisons of the simulation results between the different simulating distributions (*Uniform*, *Beta* and *Normal*) are unnecessary, we can briefly investigate their behavior. The inter-cluster variability estimates using the normal distribution to generate the simulation template were $\hat{\sigma}^2_{IC}=0.102$ for the MS approach and $\hat{\sigma}^2_{IC}=0.100$ for the MVN approach. If these levels are deemed too large, then the variance assumed in the simulation template (here $\sigma=0.1$) can be lowered. Likewise, the inter-cluster variability can be increased or decreased using the Uniform distribution by either increasing or decreasing the range about the desired proportions. While the process for the Beta distribution requires solving one equation for two unknowns (such that the given scale and shape parameters provided a desired mode), their sum can be increased or decreased to either decrease or increase the desired intra-cluster variability.

### Case 2: Clustered, Two-Group, Repeated-Measure Study

In this case we simulate $k=2$ binary outcomes over $m=20$ clusters, where subjects in half the clusters belong to a treatment group and where subjects in the other half of the clusters belong to a control group. Assuming an effective treatment, the global marginal probabilities for the two outcomes in the treatment group are $p_{11}=0.25$ and $p_{12}=0.45$, while for an ineffective control the global marginal probabilities are $p_{21}=p_{22}=0.25$; these values indicate that the difference in the changes over "time" is $(p_{12}-p_{11})-(p_{22}-p_{21})=0.20$. To incorporate inter-cluster variance we simulate the marginal probabilities according to the specifications listed in Table 4. As in the previous case, we can easily show that that each distribution obtains the target marginal probability for that Group and time. The inter-cluster variability's are similar to what was described before. The intended model is fit with a fixed two-level "time" effect, a fixed two-level "group" effect, a group-time interaction, a cluster-level random effect to account for the inter-cluster variation, and a subject-level random effect to account for the correlation between the measures. The null hypothesis is no difference in change over time between the two groups (i.e. $H_0$*:*$(p_{12}-p_{11})=(p_{22}-p_{21})$) against a two-sided alternative.

The aggregate results over the *500* simulations for case 2 are found in Table 5. Here we see again that both approaches were effective in estimating the marginal means as well as the desired difference in the change in proportions over time *(δ=0.2)*, and the efficiencies of these estimates were similar for both methods. The estimated inter-cluster variance and the correlation between the repeated measure outcomes were similar between the MS and MVN approaches, while the estimated correlations were also close to the desired level *(ρ=0.2)*. At least 98.8%of the data sets generated by the MS approach allowed models to converge, and at least 99.0% of the MVN-derived data sets allowed model convergence. The empirical powers for the testing the null hypothesis of no difference in change over time between the two groups for the multinomial approach were >99.9% (Uniform), 99.2% (Beta) and 96.8% (Normal), and were >99.9% (Uniform), 99.2% (Beta) and 96.6% (Normal) for the multivariate normal approach (not shown in Table 5).

**Table 4:** Simulation template for case two.

| | | Marginal | Distribution | | |
|---|---|---|---|---|---|
| Group | Time | Mean | Uniform | Beta | Normal |
| Treatment | 1 | $P_{11}$ | *U[0.15,0.35]* | *Beta(11,31)* | *N(0.25,0.1²)* |
| | 2 | $P_{12}$ | *U[0.35,0.55]* | *Beta(10,12)* | *N(0.45,0.1²)* |
| Control | 1 | $P_{21}$ | *U[0.15,0.35]* | *Beta(11,31)* | *N(0.25,0.1²)* |
| | 2 | $P_{22}$ | *U[0.15,0.35]* | *Beta(11,31)* | *N(0.25,0.1²)* |

## Conclusion

We extended both the multinomial sampling approach and the multivariate normal approaches to simulating dependent binary data to account for desired random effect structures. The extensions for both methods are simple to implement and offer control of marginal probabilities, dependence between outcomes, and intra-cluster variability. Rather than being assigned constant values, the desired marginal probabilities are sampled from specified probability distributions, where a separate simulation template is simulated for each cluster. Simulation studies show that our extension to both approaches yields data that achieve the desired marginal probabilities with relatively low variability, and also exhibits the desired correlation between the binary repeated measures. The parameters for the distributions used in the simulation template can also be adjusted to achieve a desired inter-cluster variability.

One limitation in the presentation of this research is that the simulation templates used here are not exhaustive. In both examples offered we only considered cases of two repeated measures and we presented a limited selection of marginal means and correlations. However, extending this approach to account for more repeated measures or alternative simulation templates is straightforward. We also did not consider more complicated random effect structures, though the underlying principle remains the same: randomly generate a simulation template for each cluster or combination of clusters. This general idea can be applied to other simulation approaches for simulating dependent binary data (e.g., Qaqish [9]), and in principle can be adapted in simulation methodologies for other types of dependent outcomes.

An important statistical role in the preparation of clustered study designs is determining the sample size required to find a desired effect size. While equations or numerical procedures for estimating

**Table 5:** Simulation results for case two.

| | Uniform | | Beta | | Normal | |
|---|---|---|---|---|---|---|
| | MS Approach | MVN Approach | MS Approach | MVN Approach | MS Approach | MVN Approach |
| $P_{11}$ | 0.247 (0.022) | 0.246 (0.023) | 0.260 (0.026) | 0.259 (0.024) | 0.248 (0.036) | 0.246 (0.034) |
| $P_{12}$ | 0.450 (0.024) | 0.452 (0.023) | 0.452 (0.036) | 0.452 (0.037) | 0.449 (0.036) | 0.450 (0.035) |
| $P_{21}$ | 0.248 (0.023) | 0.247 (0.022) | 0.259 (0.025) | 0.260 (0.026) | 0.243 (0.035) | 0.244 (0.035) |
| $P_{22}$ | 0.249 (0.024) | 0.248 (0.023) | 0.261 (0.025) | 0.258 (0.026) | 0.245 (0.034) | 0.246 (0.033) |
| $(P_{12}- P_{11})-( P_{22}- P_{21})$ | 0.203 (0.044) | 0.204 (0.044) | 0.190 (0.054) | 0.195 (0.058) | 0.199 (0.065) | 0.200 (0.066) |
| *Cluster R.E.* | 0.034 (0.020) | 0.034 (0.020) | 0.060 (0.032) | 0.063 (0.030) | 0.115 (0.055) | 0.114 (0.057) |
| $\hat{\rho}$ | 0.192 (0.023) | 0.192 (0.023) | 0.186 (0.025) | 0.183 (0.024) | 0.169 (0.026) | 0.167 (0.025) |
| % Conv. | 98.8% | 100% | 99.8% | 100% | 100% | 100% |

a required sample size are available for some situations (e.g., Donner, Birkett and Buck [10]), more complicated situations involving repeated measures and intricate clustering may require a simulation-based approach. Simulation templates can be designed to match the desired effect size and clustering structure, and empirical power can be estimated by repeatedly simulating such data. In a similar manner, new statistical methodologies suitable for repeated binary outcomes in clustered settings can be numerically assessed and compared with alternative procedures. Data can be simulated from a desired template and analyzed by the methodologies under consideration, and key features from that analysis (e.g., means, test statistics, confidence intervals, and hypothesis testing decisions) can be aggregated over repeated simulations and compared between competing models.

## Acknowledgement

The authors are thankful for the kindness and guidance offered by Dr. Karl Peace in the conduct of this research and writing of this manuscript.

### References

1. Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. The American Statistician. 1991; 45: 302-304.
2. Kang SH, Jung SH. Generating correlated binary variables with complete specification of the joint distribution. Biometrical Journal. 2001; 43: 263-269.
3. Sabo RT, Haynes ME, Chaganty NR. Using odds ratios to simulate dependent binary outcomes. Communications in Statistics: Simulation and Computation. 2015.
4. Murray DM, Perry CL, Griffin G, Harty KC, Jacobs DR, Schmid L, et al. Results from a statewide approach to adolescent tobacco use prevention. Prev Med. 1992; 21: 449-472.
5. Krist AH, Aycock RA, Etz RS, Devoe JE, Sabo RT, Williams R, et al. My Preventive Care: implementation and dissemination of an interactive preventive health record in three practice-based research networks serving disadvantaged patients. Implementation Science. 2014; 9: 181.
6. Joe H. Multivariate Models and Dependence Concepts. London: Chapman and Hall. 1997.
7. Chaganty NR, Joe H. Range of correlation matrices for dependent Bernoulli random variables. Biometrika. 2006; 93: 197-206.
8. Haynes ME, Sabo RT, Chaganty NR. Simulating dependent binary variables through multinomial sampling. Journal of Statistical Computation and Simulation. 2015.
9. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika. 2003; 90: 455-463.
10. Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. Am J Epidemiol. 1981; 114: 906-914.




**Citation:** Wang A and Sabo RT. Simulating Clustered and Dependent Binary Variables. Austin Biom and Biostat. 2015;2(2): 1020.