Underestimation of Variance Due to Using Simple Random Sampling Statistical Methods Applied to Data Obtained from a Cluster Sampling Design

Jianmin Wang, PhD; Amanda Allshouse, MS; William Irish, PhD

RTI Health Solutions, RTI International, Research Triangle Park, NC, USA

ABSTRACT

Background: A common objective of pharmaco-epidemiology is to estimate the proportion of individuals with a particular health-related issue in a well-defined target population. Two sampling strategies for this kind of objective are: simple random sampling (SRS), and cluster sampling (CS). In SRS patients are directly selected for inclusion in the sample. In CS, individual respondents are naturally aggregated into mutually exclusive and exhaustive subgroups, i.e., clusters, a sample of clusters is selected, and then one-stage all units within the cluster are included in the sample. CS is often more convenient and less expensive than SRS; however CS can result in higher variance due to within cluster homogeneity. In practice researchers frequently use variance estimation methods based on SRS when data was obtained through a CS design.

Objectives: The objectives of this study are to 1) evaluate the statistical bias inherent when using SRS methods to analyze CS data, 2) explore the relationship between design effect ([*deff*] a function of average cluster sample size and intra-cluster correlation) and variance estimation, and 3) identify when SRS statistical methods have minimal consequences when a CS design is employed.

Methods: Monte Carlo (MC) simulated cluster samples (MC=1,000) were generated under population characteristics with varying *deff*. For each simulation, variance was estimated based on SRS and CS methods. Statistical bias was evaluated by comparing the average of the MC sample variance estimates to the empirical estimate of the true variance. The design effect and the variance were plotted and the relationship was evaluated. Scenarios resulting in minimal consequences when a CS design is employed on SRS data were identified.

Results: When a SRS method is used with a CS design, the variance is underestimated. The amount by which variance is underestimated is positively correlated with *deff*; when *deff* is low the level of variance underestimation is minimal.

Conclusions: When using SRS statistical methods on data from a CS design with the large cluster size and/or high intra-cluster correlation, the precision of statistical inference will be exaggerated tremendously. However, this situation will be minimized when the average cluster size is small with low intra-cluster correlation.

CONFLICT OF INTEREST STATEMENT

This study was conducted and fully funded by RTI Health Solutions.

BACKGROUND

A common objective of pharmacoepidemiology is to estimate the proportion of individuals with a particular health-related issue in a well-defined target population.

Commonly used sampling strategies include the following:

- **Simple Random Sampling (SRS)**: Patients are directly selected for inclusion in the sample.
- **Cluster Sampling (CS)**: Individual respondents are naturally aggregated into mutually exclusive and exhaustive subgroups (i.e., clusters); a sample of clusters is selected; and then either (in a one-stage design) all units within the selected cluster are included in the sample or (in a two-stage design) some of the units within the cluster are selected randomly for inclusion in the sample.

CS is often more convenient and less expensive than SRS, but CS can result in higher variance in the prevalence estimate due to inter-cluster heterogeneity.

Methods to estimate the variance that ignore the cluster design can result in estimates that underestimate the true variance. Variance underestimation leads to narrower confidence intervals and thus overstating the results of the study. In practice, however, researchers frequently use variance estimation methods based on SRS when data was obtained using CS.

OBJECTIVES

1. Evaluate the statistical bias inherent when using SRS methods to analyze CS data
2. Explore the relationship between design effect (*deff*; a function of average cluster sample size and intra-cluster correlation) and variance estimation
3. Identify when SRS statistical methods have minimal consequences when a CS design is employed

METHODS

To evaluate the objectives, we first provide a mathematical argument for the variance and *deff* when CS is used. Next, we provide the methods used to estimate the variance and *deff* under SRS and CS strategies. Finally, we use Monte Carlo simulation to verify the statistical methods.

Theory

Table 1. Definition of Symbols

π_i	Cluster-specific prevalence	v	Cluster specific variance, i.e., variance of π_i	n_i	Cluster size
π	Overall prevalence			n	Average cluster size
$\hat{\pi}$	Overall prevalence estimate	V_{CS}	Variance of π using CS	*deff*	Design effect
i	Cluster index (i=1, 2, … k)	V_{SRS}	Variance of π using SRS	k	Number of clusters

Suppose that the clusters are independent and each cluster has exactly n units. For cluster i, the number of units with positive property X_i follows a binomial distribution with probability π_i. Because of different positive probabilities among the clusters, we assume that the cluster-specific prevalence π_i is a random variable which can take on values between 0 and 1 with mean π and variance v, $0< v <1/12$. Based on these conditions, we have following:

1. Conditional expectation of X_i for given π_i is $E(X_i \mid \pi_i) = n\pi_i$
2. Conditional variance of X_i for given π_i is $V(X_i \mid \pi_i) = n\pi_i(1-\pi_i)$
3. Marginal expectation of X_i is $E(X_i) = E[E(X_i \mid \pi_i)] = E[n\pi_i] = n\pi_i$
4. Marginal variance of X_i is
$$V(X_i) = V[E(X_i \mid \pi_i)] + E[V(X_i \mid \pi_i)] = n^2 v + n(\pi - v - \pi^2) = n(n-1)v + n\pi(1-\pi)$$

The above conditions are used to derive the variance of the prevalence estimate, which is defined as

$$\hat{\pi} = \frac{\sum_{i=1}^{k} X_i}{nk}, \qquad (1)$$

where k is the number of clusters in the sample.

The variance of prevalence estimate based on CS is

$$V_{CS} = V\left(\frac{\sum_{i=1}^{k} X_i}{nk}\right) = V\left(\sum_{i=1}^{k} X_i\right) \Big/ n^2k^2 = V(X_i) \Big/ n^2 k = \frac{n(n-1)v + n\pi(1-\pi)}{n^2 k} = \frac{(n-1)v}{nk} + \frac{\pi(1-\pi)}{nk}$$

If SRS is used, then the variance of the prevalence estimate is

$$V_{SRS} = \frac{\pi(1-\pi)}{nk}$$

The *deff* is defined as the ratio of the two sampling variances:

$$deff = \frac{V_{CS}}{V_{SRS}} = 1 + (n-1)\frac{v}{\pi(1-\pi)}, \; deff \geq 1 \qquad (2)$$

This formula indicates:

1. Prevalence estimation based on CS has large variation compared to estimation based on SRS
2. When cluster size = 1 or v = 0, the CS design is equivalent to SRS
3. Overall prevalence π, variance of the cluster-specific prevalence v, and cluster size n affect the *deff*.

Figure 1 illustrates the probability density function (pdf) for a random variable π_i following a beta distribution with the expected prevalence fixed at 25% and v ranging from 0.001 to 0.02, where wider ranges have a flatter density function. Figure 2 depicts the simple linear relationship between *deff* and cluster size based on the pdfs in Figure 1 when π = 25% and v ranges from 0.001 to 0.02. For example, if the range of cluster-specific prevalence is between 0.18 and 0.34, as seen in Figure 1 in the most peaked distribution, then in Figure 2 the *deff* would be linearly related to cluster size with v = 0.001 and the variance inflation factor would be minimally impacted. The *deff* and cluster size always have a simple linear relationship for any given π and v.

Estimation

When SRS is used, variance of the prevalence estimate is estimated as

$$\hat{V}_{SRS} = \frac{\hat{\pi}(1-\hat{\pi})}{\sum_{i=1}^{k} n_i} \qquad (3)$$

When CS is used, variance of the prevalence estimate is estimated as

$$\hat{V}_{CS} = \frac{k}{(k-1)}\left(\sum_{i=1}^{k} X_i^2 + \hat{\pi}^2 \sum_{i=1}^{k} n_i^2 - 2\hat{\pi}\sum_{i=1}^{k} n_i X_i\right)\left(\sum_{i=1}^{k} n_i\right)^{-2} \qquad (4)$$

(Kish, 1965).

Monte Carlo Simulation

Monte Carlo Simulation was used to verify the statistical method used for estimation (i.e., SRS or CS). Data were simulated for the different population characteristics using the following steps:

1. Set population parameters (i.e., overall prevalence π, variance of cluster-specific prevalence v, number of clusters k, and average cluster size n).
2. Assume that cluster size follows uniform distribution bounded by 0.85 × n to 1.15 × n.
3. Generate data according to specified probability distribution.
4. Estimate the overall prevalence π and its variance v by SRS and CS methods described above.
5. Repeat steps (3) and (4) 1000 times.
6. Calculate the empirical mean estimate of the SRS and CS variance and *deff*.

Figure 1. Density functions of cluster-specific prevalence

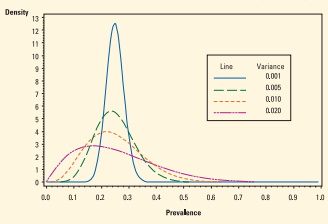
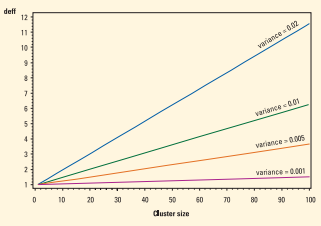


Figure 2. Relationship between *deff*, *n*, and *v* when π = 25% and *v* ranges from 0.001 to 0.02



RESULTS

Bias:

- $\hat{\pi}$ is an unbiased estimate of the overall prevalence π. The difference between $\bar{\hat{\pi}}$ and π is very near zero (Table 2).
- $\hat{V}_{CS}$ and $\hat{V}_{SRS}$ are unbiased estimates of V_{CS} and V_{SRS}, respectively. The difference between the empirical-based estimator and the theoretical results are very near zero (Table 2).

Relationship between *deff* and variance estimation:

- Number of clusters, k does not impact *deff*. Theoretical and empirical *deffs* are similar and are constant for any given number of clusters (Figure 3).
- For fixed overall prevalence, *deff* increases with increasing cluster size and increasing cluster-specific variance (Figures 4 to 6).
- For fixed cluster size and cluster-specific variance, *deff* decreases with increasing overall prevalence when overall prevalence is less than 0.50 (Figures 4 to 6) and increases with increasing overall prevalence when overall prevalence is greater than 0.50 (data not shown).

Minimizing consequences from using SRS methods when a CS design is employed:

- The empirical *deff* and theoretical *deff* displayed in Figures 3 to 6 are nearly identical, which validates formula 2. As a result, we can derive the relationship between bias and *deff*, which is applicable to any sampling design.
$$\text{Bias} = V_{SRS} - V_{CS} = V_{CS} * (1\text{-}deff)/deff$$
- Bias will be smallest when *deff* is close to one (i.e., average cluster size is close to one, or variance of the cluster-specific prevalence is close to zero).

Table 2. *Deff* Based on Theoretical and Empirical as Function of the Number of Clusters and Cluster Size

Population Parameters*				Theoretical			Empirical**				
k	n	π	v	V_{SRS}	V_{CS}	*deff*	$\bar{\hat{\pi}}$	$V(\hat{\pi})$	$\bar{V}_{CS}$	$\bar{V}_{SRS}$	*deff*
20	20	0.1	0.015	0.2250	0.9375	4.1667	0.0999	0.8846	0.9509	0.2284	4.1628
50	80	0.05	0.001	0.0119	0.0316	2.6632	0.0499	0.0335	0.0317	0.0119	2.6575
50	30	0.05	0.01	0.0317	0.225	7.1053	0.0498	0.2334	0.2251	0.0319	7.0492
50	30	0.05	0.02	0.0317	0.4183	13.211	0.05	0.4303	0.4202	0.0319	13.165
50	100	0.05	0.02	0.0095	0.4055	42.684	0.0494	0.3926	0.404	0.0094	43.138
50	30	0.25	0.001	0.1250	0.1443	1.1547	0.2495	0.1466	0.1465	0.1268	1.1551
50	35	0.25	0.01	0.1071	0.3014	2.8133	0.2505	0.3007	0.3047	0.1087	2.8027
50	70	0.25	0.01	0.0536	0.2507	4.68	0.2504	0.2638	0.2527	0.054	4.6816

Theoretical and empirical variance are displayed as the value times 10^{-3}

*Only eight out of approximately 300 combinations of k, n, π, and v are presented in the table

**Additional formulas for statistics presented in table are given below:

$$\bar{\hat{\pi}} = \sum_{j=1}^{1000} \hat{\pi}_j / 1000 \qquad \bar{V}_{CS} = \sum_{j=1}^{1000} \hat{V}_{CSj} / 1000 \qquad \bar{V}_{SRS} = \sum_{j=1}^{1000} \hat{V}_{SRSj} / 1000 \qquad V(\hat{\pi}) = \sum_{j=1}^{1000} (\hat{\pi}_j - \bar{\hat{\pi}})^2 / 999$$

Figure 3. Relationship between *deff* and number of clusters

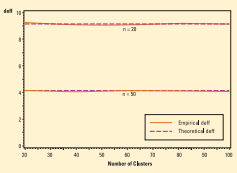


Figure 4. Empirical and theoretical *deff* given overall prevalence being 0.05

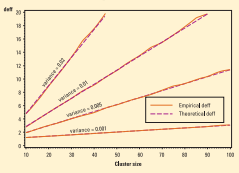


Figure 5. Empirical and theoretical *deff* given overall prevalence being 0.15

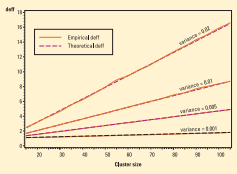
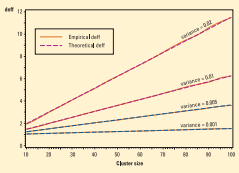


Figure 6. Empirical and theoretical *deff* given overall prevalence being 0.25



CONCLUSIONS

- Using SRS statistical methods to estimate the variance of the overall prevalence when data is collected under a CS design will result in variance estimates that underestimate the true variance, especially when the average cluster size and/or intra-cluster correlation is large.
- If the average cluster size is small (i.e., $n \leq 5$) or intra-cluster correlation is small (i.e., v < 0.001), then SRS methods will have minimal impact.
- However, we recommend using formula 4 to estimate variance when a CS design is used.

REFERENCE

Kish, L: *Survey Sampling*. New York: John Wiley & Sons; 1965.

CONTACT INFORMATION

Jianmin Wang, PhD
Director, Biostatistics

RTI Health Solutions, RTI International
3040 Cornwallis Road, PO Box 12194
Research Triangle Park, NC 27709-2194

Phone: 919-541-6986
Fax: 919-541-7222
E-mail: jmwang@rti.org

Presented at: 22nd International Conference on Pharmacoepidemiology and Therapeutic Risk Management
August 24–27, 2006
Lisbon, Portugal