Published in final edited form as:

Stat Med. 2018 June 15; 37(13): 2174–2186. doi:10.1002/sim.7648.

Grouping methods for estimating the prevalences of rare traits from complex survey data that preserve confidentiality of respondents

Noorie Hyun¹, Joseph L. Gastwirth², Barry I. Graubard³

¹Division of Biostatistics, Institute of Health and Equity, Medical College of Wisconsin, Milwaukee, WI, USA

²Department of Statistics, George Washington University, Washington, DC, USA

³Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, U.S.A.

Abstract

Originally, 2-stage group testing was developed for efficiently screening individuals for a disease. In response to the HIV/AIDS epidemic, 1-stage group testing was adopted for estimating prevalences of a single or multiple traits from testing groups of size q , so individuals were not tested. This paper extends the methodology of 1-stage group testing to surveys with sample weighted complex multistage-cluster designs. Sample weighted-generalized estimating equations are used to estimate the prevalences of categorical traits while accounting for the error rates inherent in the tests. Two difficulties arise when using group testing in complex samples: (1) How does one weight the results of the test on each group as the sample weights will differ among observations in the same group. Further-more, if the sample weights are related to positivity of the diagnostic test, then group-level weighting is needed to reduce bias in the prevalence estimation; (2) How does one form groups that will allow accurate estimation of the standard errors of prevalence estimates under multistage-cluster sampling allowing for intracluster correlation of the test results. We study 5 different grouping methods to address the weighting and cluster sampling aspects of complex designed samples. Finite sample properties of the estimators of prevalences, variances, and confidence interval coverage for these grouping methods are studied using simulations. National Health and Nutrition Examination Survey data are used to illustrate the methods.

Keywords

categorical outcome; group testing; NHANES; probability sample; sensitivity and specificity

Correspondence Noorie Hyun, Division of Biostatistics, Institute of Health and Equity, Medical College of Wisconsin, Milwaukee, WI, USA. nhyun@mcw.edu.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

1 | INTRODUCTION

Group testing was originally proposed by Dorfman¹ for efficiently screening a population to identify individuals that have a rare binary trait (disease). At the first stage, the samples of individuals are combined into groups of size q , and each group is tested. If a group tests positive, then every member of the group is retested to identify the individuals in the group who have positive test results. Later researchers have refined the methodology. For example, using more stages of groupings by hierarchically subdividing the initial groups that test positive into smaller groups to subsequently identify positive subgroups,^{2,3} using partial hierarchical subdivision called halving,⁴ or using arrays to group samples (see Barillot et al⁵ and Phatarford and Sudbury,⁶ for application to genetic testing), which can be competitive with hierarchical grouping.⁷ Other issues related to group testing have been studied such as dilution effects from the grouping of the samples^{8,9} and misclassification because of imprecise trait identification.¹⁰

In response to the HIV/AIDS epidemic, group testing procedures of biological specimens that do not test individuals were developed by Cahoon-Young et al¹¹ and Gastwirth and Hammick¹² to estimate the prevalence of infections while preserving the privacy of the individual participants. Further refinements of these procedures were proposed for more prevalent traits¹³ and for simultaneously estimating 2 or more traits.¹⁴ Maintaining confidentiality is a major concern for study participants from national surveys such as the National Health and Nutrition Examination Survey (NHANES) that collect for trait determination, eg, see the National Center for Health Statistics consent forms and confidentiality brochure.¹⁵ These surveys use complex sample designs with stratified multistage-cluster sample selection that causes individuals to be sampled with differing probabilities. Because the probabilities can be related to the prevalence of the trait (ie, informative sampling), then weighting observations is required to obtain approximately unbiased estimates of the population parameters. Also, variance estimation needs to account for potential correlation among observations from individuals sampled from the same clusters because they can have characteristics and measurements that are more similar than individuals across clusters (ie, positive intracluster correlation). The standard methods for estimation from simple random samples will underestimate the sampling error because this positive correlation inflates the variances of estimates. This paper considers statistical methods that address the effect of complex sample designs on prevalence estimates obtained from group testing.

Previous research in group testing with data from complex sample designs has focused on continuous measured data for the purpose of reducing the number of samples that are below limits of laboratory detection, eg, pesticide exposure.^{16,17} In that research, the grouping of the samples was entirely random so that the groups included individuals from different sampling strata and clusters. Because unbiased variance estimation from multistage stratified cluster samples requires that the observations can be partitioned by the sampling strata and clusters used in the first-stage of sampling, there was no direct method to produce unbiased variance estimates for the results obtained. In this paper, we consider 5 different grouping methods to address the sample weighting and cluster sampling aspects of complex designed samples and allow for sample design-based estimation. The proposed grouping methods and

estimation procedures yield approximately unbiased prevalence estimates variances for the finite population.

Section 2 reviews prevalence estimation methods under group testing for a single trait and for 2 traits for simple random samples and introduces prevalence estimation for group testing with complex sample designs. We also account for misclassification of test measurements by incorporating the sensitivity and specificity of the test in the likelihoods (and pseudo-likelihoods for complex samples) for prevalences. In Section 3, statistical properties of the estimators of prevalences for the 5 grouping methods are compared using simulation studies that generate data reflecting the complex sample design of the NHANES. In Section 4, we applied the proposed grouping methods to assessing oral Human Papillomavirus (HPV) deoxyribonucleic acid (DNA) types from the NHANES 2009 to 2012 data. In Section 5, we discuss implications of the simulation results and application and conclude with recommendations.

2 | METHODS

2.1 | Estimation of the prevalence of 1 trait from a simple random sample

Suppose we test for a trait X in a simple random sample of n subjects selected without replacement from a large finite population of size N (a constant), where $X_i = 1$ if the screening test on subject i is positive, $X_i = 0$ if negative. We allow for the possibility of misclassification in the testing of the trait by letting η and θ denote the sensitivity and specificity of the test, which are defined as $Pr(X_i = 1 | D_i = 1)$ and $Pr(X_i = 0 | D_i = 0)$, respectively, where $D_i = 1$ if subject i has the trait, $D_i = 0$ otherwise for $i = 1, \dots, n$. We can express the probability of a positive test with η , θ , and the prevalence $\pi = Pr(D_i = 1)$ as

$$\begin{aligned} Pr(X_i = 1) &= Pr(D_i = 1, X_i = 1) + Pr(D_i = 0, X_i = 1) \\ &= Pr(X_i = 1 | D_i = 1)Pr(D_i = 1) + Pr(X_i = 1 | D_i = 0)Pr(D_i = 0) \\ &= \eta\pi + (1 - \theta)(1 - \pi), \end{aligned} \quad (1)$$

where $0 \leq \eta, \theta, \pi \leq 1$ and η and θ are assumed to be known and π is an unknown parameter.

For grouped data with a group size of q , there are $n/q = m$ groups, instead of observing the X_i , we observe $X_g = 1$ if at least 1 sample in group g has a positive test result, and $X_g = 0$ otherwise for $g = 1, \dots, m$. The probability of group g testing positive is $Pr(D_g = 1) = 1 - (1 - \pi)^q$, where $D_g = 1$ if at least 1 sample in group g has the trait, $D_g = 0$ otherwise. The accuracy of grouped tests can be different from the ungrouped test. Given the values of $\eta^* = Pr(X_g = 1 | D_g = 1)$ and $\theta^* = Pr(X_g = 0 | D_g = 0)$,

$$\begin{aligned} Pr(X_g = 1) &= Pr(X_g = 1 | D_g = 1)Pr(D_g = 1) + Pr(X_g = 1 | D_g = 0)Pr(D_g = 0) \\ &= \eta^* \{1 - (1 - \pi)^q\} + (1 - \theta^*)(1 - \pi)^q \\ &:= \psi(\pi; \eta^*, \theta^*). \end{aligned} \quad (2)$$

Since we assume for $g = 1, \dots, m$, each X_g follows a Bernoulli distribution with probability $\psi(\pi; \eta^*, \theta^*)$ of a positive test, and X_g is mutually independent, we can apply standard maximum-likelihood methods to obtain point and variance estimation for a finite population. In the following sections, we adopt weighted-likelihood methods for multistage stratified cluster sample with differential sampling probabilities.

2.2 | Estimation of the prevalence of 1 trait with a multistage stratified cluster sample

We consider a target finite population of individuals that is partitioned into J primary sampling units (PSUs) usually determined by geographic location of residence, eg, counties or contiguous counties, in national household surveys. The PSUs are placed into L strata with J_l PSUs in stratum $l = 1, \dots, L$ (the L sampling strata are disjoint and exhaustive). Let N_{lj} be the number of individuals in the population that are in PSU j in stratum l . At the first stage of sampling, a random sample of k_l PSU(s) is selected without replacement from each of the strata with probabilities proportional to a measure of size, eg, N_{lj} . At the second and later stages, stratification and cluster sampling often are used to select a random sample of n_{lj} individuals from the N_{lj} individuals without replacement within the sample PSUs (Figure 1). For notational simplification, we assume that the number of PSUs that are sampled in each stratum is constant $k \geq 2$. Let $w_{l,j,i}$ denote sample weight for an individual in the sample, which is the reciprocal of the product of the probabilities of sampling the PSU j in stratum l and the probability of sampling individual i in sampled PSU j in stratum l (Korn and Graubard¹⁸, p. 20).

To account for the intracluster correlation within PSUs in the designed-based variance estimation of the estimation of the prevalences that is presented in section 2.2.1, groups are formed from the observations within each sampled PSU. We base our estimation of π on the following (pseudo) weighted log-likelihood of the observations

$$\{X_{l,j,g}; l = 1, \dots, L, j = 1, \dots, k, g = 1, \dots, m_{l,j}\}:$$

$$\sum_{l,j,g} w_{l,j,g} [X_{l,j,g} \log \psi(\pi; \theta^*, \eta^*) + (1 - X_{l,j,g}) \log \{1 - \psi(\pi; \theta^*, \eta^*)\}]. \quad (3)$$

We need to calculate what the joint probability of sampling and grouping together the q subjects; then an accurate value of $w_{l,j,g}$ could be used in (3) resulting in consistent estimate of π (assuming the π was the same across PSUs). However, for complex samples, it is an intractable to compute these joint inclusion probabilities. Instead, we follow Caudill¹⁷ using the average of the individual sampling weights of the subjects sampled and grouped together, that is, $\sum_{i \in g} w_{l,j,i} / q = \bar{w}_{l,j,g}$.

There are 2 sources of bias when estimating the prevalence π using (3), where the grouping is nested within the sampled PSUs. The first occurs when the prevalences vary across the PSUs as would be the case when the intracluster correlation is nonzero because the groups formed from individuals in the same PSU do not correspond to the groups randomly formed from the entire population. This results in a potential bias. However, the nesting of groups within PSUs is needed to allow for design-based estimation of variances for the estimated π .

The second source arises from weighting the group testing results by the average group-level weight $\bar{w}_{l,j,g}$ for the q subjects when the sample weights are informative, ie, they are related to the probability that a sampled individual's test is positive. To reduce the bias from using the group average weight, we consider grouping samples that have similar sample weights. Unfortunately, these groups will differ from random groups from the entire sample, which could further increase the bias. This effect will be evaluated in our simulations. It is worth noting that despite the bias, an advantage of using the group average weight is that the estimated π will have smaller variance than the ungrouped estimated π because the average group-weights will be less variable than the individual-level weights.

We consider 5 grouping methods within each cluster in order to obtain less biased estimation using average weight (Figure 2): method 1 (M1), randomly group observations; method 2 (M2), sort observations by their sample weights and group q adjacent observations; method 3 (M3), group the sorted observations by their sampling weights in a serpentine fashion (that is, we assign each of the sorted observations to group, $g = 1, 2, \dots, m-1, m, m, m-1, \dots, 2, 1, 1, 2, \dots$); method 4 (M4), first stratify the observations into high/low sampling weight strata, and then randomly form groups of observations within strata; method 5 (M5), the same as M4 but stratifying into 3 categories, low/medium/high. Grouping methods 2 to 5 are designed to make the sampling weights of the grouped observations more homogeneous than the random grouping method 1.

The following 2 paragraphs are heuristic and are meant to provide the reader with a sense of how the bias occurs for the group testing methods, described here, in complex samples. Let the finite population consist of all the PSUs and all the individuals in the PSUs. Let the individuals be divided up into groups of size q (or at least approximately q), where the groups were nested in the PSUs. Let the probability of selecting each group in the sample be the reciprocal of the group average weight then the weighted log-likelihood (3) is estimating the log-likelihood of $X_{i,j,g}$ for all the groups in the finite population. However, because the grouping is done after the selection of the sample of individuals, the equation (3) for the sample is an approximation to the previously mentioned log-likelihood for the finite population. If the sampling design is ignorable, that is, the sampling weighting is ignorable and the intracluster correlation is zero, then maximizing (3) will produce consistent estimates of π . If the sampling design is not ignorable, then the inference needs to take it into account.

For approximate consistency of the estimation, a sufficient condition is

$$E_1 \left[E_2 \left\{ \sum_{l=1}^L \sum_{j=1}^k \sum_{g=1}^{m_{l,j}} \bar{w}_{l,j,g} f(X_{l,j,g}; \pi) \right\} \right] \approx E_1 \left\{ \sum_{l=1}^L \sum_{j=1}^{J_l} \sum_{g=1}^{m_{l,j}} f(X_{l,j,g}; \pi) \right\} \approx \sum_g^G f(X_g; \pi), \quad (4)$$

where $f(X_{l,j,g}; \pi) = X_{l,j,g} \log \psi(\pi; \theta^*, \eta^*) + (1 - X_{l,j,g}) \log \{1 - \psi(\pi; \theta^*, \eta^*)\}$.

E_2 is the expectation over repeated sampling from the finite population; E_1 is the expectation over repeatedly grouping samples in the finite population using the same grouping method

as used in sample; and the grouping in the last term of (4) is across strata and PSUs. The first approximation in (4) indicates the expectation of E_2 in the first term of (4) converges under appropriate conditions of increasing sample and finite population sizes. We expect that the first approximation in (4) will hold under sample designs without highly informative weighting with respect the binary trait or group size q is small. We expect that the second approximation will hold when the prevalence of the trait does not vary greatly across the PSUs.

2.2.1 | Inference procedure—After reparametrizing π by $\exp(\alpha)/\{1 + \exp(\alpha)\}$ to guarantee a positive estimate for π and assuming that η^* and θ^* are known, the weighted estimating equations for α are obtained by differentiating the weighted log-likelihood (3) with respect to α to obtain

$$U_n(\alpha) = \sum_{l=1}^L \sum_{j=1}^k \sum_{g=1}^{m_{l,j}} \bar{w}_{l,j,g} U(\alpha; X_{l,j,g}) = 0, \quad (5)$$

$$\text{where } U(\alpha; X_{l,j,g}) = \left[\frac{X_{l,j,g}}{\psi(\alpha)\{1 - \psi(\alpha)\}} - \frac{1}{1 - \psi(\alpha)} \right] \frac{d\psi(\alpha)}{d\alpha}; \quad (6)$$

$\psi(\alpha) = \eta^*[1 - \{1 - \pi(\alpha)\}^q] + \{1 - \theta^*\}(1 - \pi(\alpha))^q$;
 $d\psi(\alpha)/d\alpha = q\exp(\alpha)\{1 + \exp(\alpha)\}^{-q-1}(\eta^* + \theta^* - 1)$; $m_{l,j} = n_{l,j}/q$. This equation (6) is regarded as a weighted-generalized estimating equation using an independent working correlation.

The estimator $\hat{\alpha}$ such that $U_n(\hat{\alpha}) = 0$ can be obtained by the Newton-Raphson method, and the estimator asymptotically converges to a normal distribution (the regularity conditions are summarized in Section 1 in the Supplementary Material):

$$\sqrt{n}(\hat{\alpha} - \alpha^*) = \left\{ \frac{dU_n(\alpha)}{d\alpha} \right\}_{\alpha=\alpha^*}^{-1} \sum_{l=1}^L \sum_{j=1}^k \sum_{g=1}^{m_{l,j}} \bar{w}_{l,j,g} U(\alpha^*; X_{l,j,g}) + o_p(1) \rightarrow N(0, \sigma^2), \quad (7)$$

where $n = \sum_{l=1}^L \sum_{j=1}^k n_{l,j}$ and α^* is the limit of $\hat{\alpha}$ under increasing sample and population sizes. To estimate the asymptotic variance σ^2 , we use the Taylor linearization¹⁹ term in (7):

$$\hat{\sigma}^2 = n \left\{ \frac{U_n(\hat{\alpha})}{d\alpha} \right\}^{-2} \sum_{l=1}^L \frac{k}{k-1} \sum_{j=1}^k \left\{ \Delta_{l,j}(\hat{\alpha}) - \bar{\Delta}_l(\hat{\alpha}) \right\}^2, \quad (8)$$

$$\text{where } \Delta_{l,j}(\hat{\alpha}) = \sum_{g=1}^{m_{l,j}} \bar{w}_{l,j,g} U(\hat{\alpha}; X_{l,j,g}) \text{ and } \bar{\Delta}_l(\hat{\alpha}) = \sum_{j=1}^k \Delta_{l,j}(\hat{\alpha})/k.$$

The asymptotic variance for $\hat{\pi}$ can be estimated by application of the delta method using the variance estimator in (8), that is, yielding $\widehat{\text{var}}(\hat{\pi}) = \widehat{\text{var}}(\hat{\alpha}) \exp(2\hat{\alpha}) / \{1 + \exp(\hat{\alpha})\}^4$.

2.3 | Estimation of the prevalences of 2 traits with a multistage stratified cluster sample

In this section, we will follow the notation of Pfeiffer et al,¹⁴ who consider joint estimating 2 prevalences of 2 traits from random grouping of simple random samples of DNA, to estimate the prevalences of 2 traits from multistage stratified cluster samples.

We refer to the 2 traits as A and B , and let $D_{iA} = 1$ denote the event that individual i has the trait A , and $D_{iA} = 0$ otherwise and similarly define the events $D_{iB} = 1$ and $D_{iB} = 0$ for trait B . Let $\pi_{ab} = P(D_A = a, D_B = b)$ for $a, b = 0, 1$. Then, π_{00} and π_{11} denote the probability that a subject has no and both traits, respectively. Similarly, π_{10} and π_{01} denote the probability that a subject has a trait A and B , respectively.

For simplicity, we omit the subscripts for the strata and PSUs in this section. For m groups with size q , let X_{gA} be one if group g tests positive for trait A and zero otherwise; analogously, we define X_{gB} for trait B . Let D_{gA} be one if there is at least 1 subject with trait A in group g and zero otherwise; we define D_{gB} for trait B likewise. The sensitivity of the test for a given group size for trait A is defined as $\eta_A^* = Pr(X_{gA} = 1 | D_{gA} = 1)$, and the specificity is $\theta_A^* = Pr(X_{gA} = 0 | D_{gA} = 0)$. We define η_B and θ_B analogously for trait B . In addition, we assume that X_{gA} depends only on D_{gA} and X_{gB} depends only on D_{gB} , so that

$$Pr(X_{gA}, X_{gB} | D_{gA}, D_{gB}) = Pr(X_{gA} | D_{gA}) Pr(X_{gB} | D_{gB}). \quad (9)$$

We define 4 random variables: $Y_{g11} = X_{gA}X_{gB}$, indicating a positive test for both traits, $Y_{g10} = X_{gA}(1 - X_{gB})$, indicating a positive for trait A only, $Y_{g01} = (1 - X_{gA})X_{gB}$, indicating a positive for trait B only, and $Y_{g00} = 1 - (Y_{g11} + Y_{g10} + Y_{g01})$, indicating the test is negative for both traits. The vector $(Y_{g11}, Y_{g10}, Y_{g01}, Y_{g00})$ has a multinomial distribution with probabilities p_{ab} for $a, b = 0, 1$. We assume that the outcome of 2 tests are independent, ie, testing for trait A does not affect the test for trait B . The joint probabilities that at least 1 subject in group g has neither trait A nor B , trait A only, trait B only, or trait A and B are

$$Pr(D_{gA} = 0, D_{gB} = 0) = \pi_{00}^q, \quad (10)$$

$$Pr(D_{gA} = 1, D_{gB} = 0) = \sum_{i=1}^q \frac{q!}{i!(q-i)!} \pi_{10}^i \pi_{00}^{q-i} = (\pi_{10} + \pi_{00})^q - \pi_{00}^q, \quad (11)$$

$$Pr(D_{gA} = 0, D_{gB} = 1) = \sum_{i=1}^q \frac{q!}{i!(q-i)!} \pi_{01}^i \pi_{00}^{q-i} = (\pi_{01} + \pi_{00})^q - \pi_{00}^q, \quad (12)$$

$$Pr(D_{gA} = 1, D_{gB} = 1) = 1 - (\pi_{10} + \pi_{00})^q - (\pi_{01} + \pi_{00})^q + \pi_{00}^q. \quad (13)$$

The probabilities that group g tests positive for both traits, trait A only, and trait B only are expressed in (14) using the sensitivity and specificity value of $\eta_A^*, \eta_B^*, \theta_A^*, \theta_B^*$ and the joint probabilities in (10) to (13) (the details are presented in Section 2 in the Supplementary Material):

$$p_{ab} = \sum_{D_{gA}, D_{gB}} Pr(X_{gA} = a | D_{gA}) Pr(X_{gB} = b | D_{gB}) Pr(D_{gA}, D_{gB}), \quad (14)$$

for $a, b = 0$ and 1 , and $\sum_{a=0}^1 \sum_{b=0}^1 p_{ab} = 1$.

The weighted log-likelihood for the grouped observations,

$Y = \{Y_{l,j,g} = (Y_{l,j,g11}, Y_{l,j,g10}, Y_{l,j,g01}, Y_{l,j,g00}) | l = 1, \dots, L, j = 1, \dots, k, g = 1, \dots, m_{l,j}\}$ (recall there are L sampling strata, k sampled PSUs from each stratum, and $m_{l,j}$ groups formed for testing) is, when we define $\pi_{-1} = (\pi_{00}, \pi_{10}, \pi_{01})$,

$$l^w(\pi_{-1} | Y_{l,j,g}, \eta_A^*, \eta_B^*, \theta_A^*, \theta_B^*) = \bar{w}_{l,j,g} \left(\sum_{a,b=0}^1 Y_{l,j,g,ab} \log p_{ab} \right), \quad (15)$$

$$l_n^w(\pi_{-1} | Y, \eta_A^*, \eta_B^*, \theta_A^*, \theta_B^*) = \sum_{l=1}^L \sum_{j=1}^k \sum_{g=1}^{m_{l,j}} l^w(\pi_{-1} | Y_{l,j,g}), \quad (16)$$

and the estimating equation derived from the weighted likelihood in (16) is

$$U_n(\boldsymbol{\pi}_{-1}) = \sum_{l=1}^L \sum_{j=1}^k \sum_{g=1}^{m_{l,j}} U(\boldsymbol{\pi}_{-1} | Y_{l,j,g}) \quad (17)$$

where $U(\boldsymbol{\pi}_{-1} | Y_{l,j,g}) = \partial l^w(\boldsymbol{\pi}_{-1} | Y_{l,j,g}) / \partial \boldsymbol{\pi}_{-1}$. The solution of Equation 17, $\hat{\boldsymbol{\pi}} = \{\hat{\pi}_{ab}; a, b = 0 \text{ or } 1\}$, can be obtained by a standard optimization program (eg, optimx-package in R)²⁰ with linear constraints, $0 \leq \pi_{ab} \leq 1$ for $a, b = 0, 1$ and $\sum_{a,b=0}^1 \pi_{ab} = 1$. As discussed in section 2.2.1, the estimator $\hat{\boldsymbol{\pi}}_{-1} = (\hat{\pi}_{00}, \hat{\pi}_{10}, \hat{\pi}_{01})$ also converges to a multivariate normal distribution with mean $\boldsymbol{\pi}_{-1}^* = (\pi_{00}^*, \pi_{10}^*, \pi_{01}^*)$, which can be the true parameter.

Applying the Taylor linearization method, the asymptotic variance estimator for $\hat{\boldsymbol{\pi}}_{-1}$ is

$$\widehat{var}(\hat{\boldsymbol{\pi}}_{-1}) = \sum_{l=1}^L \frac{k}{k-1} \sum_{j=1}^k \left\{ \Delta_{l,j}(\hat{\boldsymbol{\pi}}_{-1}) - \bar{\Delta}_l(\hat{\boldsymbol{\pi}}_{-1}) \right\}^2, \quad (18)$$

where $\Delta_{l,j}(\hat{\boldsymbol{\pi}}_{-1}) = \left\{ \partial U_n(\boldsymbol{\pi}_{-1}) / \partial \boldsymbol{\pi}_{-1} | \boldsymbol{\pi}_{-1} = \hat{\boldsymbol{\pi}}_{-1} \right\}^{-1} \sum_{g=1}^{m_{l,j}} U(\hat{\boldsymbol{\pi}}_{-1}; Y_{l,j,g})$ and $\bar{\Delta}_l(\hat{\boldsymbol{\pi}}_{-1}) = \sum_{i=1}^k \Delta_{l,i}(\hat{\boldsymbol{\pi}}_{-1}) / k$. Since $\hat{\pi}_{11} = 1 - \hat{\pi}_{00} - \hat{\pi}_{10} - \hat{\pi}_{01}$, avariance estimator for $\hat{\pi}_{11}$ is $\mathbf{1}^T \widehat{var}(\hat{\boldsymbol{\pi}}_{-1}) \mathbf{1}$, where $\mathbf{1}$ is a constant vector of all ones.

3 | SIMULATION

We conducted a series of simulation studies to investigate the numerical performance of the 5 grouping methods for estimating a prevalence of a single trait and a prevalence of 2 traits when the sampling weights are noninformative and informative. The sampling structure of the simulation reflects the complex sample design of the NHANES,^{21,22} which is the application data. For each simulation study, we first generate a large finite population of consisting of PSUs within strata and observations for individuals within the psus. Next, we generate samples from the finite population that are stratified 2-stage cluster samples. The simulations are repeated 5000 times where the large finite population is regenerated for each repeated simulation. We assume that the sensitivity and specificity of the both ungrouped and grouped test are 80% and 95%, respectively. Table 1 shows the different simulation scenarios. The first scenario investigates the extent of bias in the estimates when the sampling weights are ignored. In the second scenario, we use the weighted-generalized estimating equations with known sensitivity and specificity values and compare the mean square errors of the prevalence estimates obtained from the 5 grouping methods with that obtained from testing individual observations, ie, no grouping. In the first and second scenarios, the sensitivity and specificity of the test are known, whereas in the third scenario, the sensitivity and specificity are unknown, and we consider cases when incorrect values of sensitivity and specificity are used. The third study reflects a common situation when

authors cite a paper reporting sensitivity and specificity values for the specific test used but are unsure whether these values are applicable to their study population.

For testing the prevalence of a single trait, we assume π is distributed as a beta-distribution across the PSUs in the finite population with 2 shape parameters; that is, the density function is $\Gamma(a+b)/(\Gamma(a)\Gamma(b))x^{(a-1)}(1-x)^{(b-1)}$, where $(a=1, b=99)$ and $(a=7, b=93)$ resulting in expected prevalences of 1% and 7%, respectively. We generate a true prevalence rate π from the beta-distribution for each PSU in the finite population so that the response Y is beta-binomial distributed with the intracluster correlation of 0.01. This induces intracluster correlation of the observations within each of the population PSUs.

For testing the prevalence of 2 traits, we assume $\pi = (\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11})$ is distributed across the PSUs as a Dirichlet distribution with density given by

$$f(\pi | \bar{\pi}, \zeta) = \Gamma(\zeta) \prod_{a,b=0,1} (\pi_{ab})^{\zeta \bar{\pi}_{ab} - 1} / \prod_{a,b=0,1} \Gamma(\zeta \bar{\pi}_{ab}), \quad (19)$$

where with the expectation $\bar{\pi} = (92.36, 3.25, 3.25, 1.14)$ and $\zeta = 0.01$ so that the intracluster correlation²³ $\rho = \zeta/(1 + \zeta) = 0.0099$.

For testing 2 traits, we generate a joint prevalence π for each PSU in the finite population according to a Dirichlet distribution (19) and assume the tests for traits A and B have the same sensitivity and specificity. The responses $Y = (Y_{00}, Y_{10}, Y_{01}, Y_{11})$ for each PSU are generated as a multinomial distribution with the true probabilities π , so the response Y is a Dirichlet-multinomial distributed. To incorporate a test's sensitivity η and specificity θ we generate observed test results separately for the true trait group and true nontrait group as binomial distributions with q and $1 - \theta$, respectively. The same η and $1 - \theta$ are used for the individual observations.

We generate a finite population consisting of 30 strata, 100 PSUs per stratum, and 1500 subjects per PSU for total population size of $N = 4\,500\,000$. At the first stage, of sampling, we sample 2 PSUs without replacement from each stratum with probability proportional to a measure of size that is known for each PSU in the population, and at the second stage, separately, we sample 150 subjects from each sampled PSU without replacement with probability proportional to another measure of size that is known for each subject in the sampled PSU. The measure of size at the first stage of sampling, z_1 , is generated from a noncentral chi-square distribution with 5 degrees of freedom and noncentrality parameter 5. Because the first stage measure of size of a PSU is unrelated to testing an individual's having the trait, the sampling weights for the first stage are noninformative. For the second stage sampling, the measure of size is generated in 2 ways to produce either noninformative or informative probabilities of selection. In the first, the measure of size is generated independently for each study subject in a sampled PSU using a random variable z_2 distributed as noncentral chi-square distribution with 5 degrees of freedom and noncentrality parameter 5, as this measure of size is unrelated to the testing of the traits, the resulting second stage sampling weights are noninformative. In the second way, the measure of size is

generated independently for each study subject in a sampled PSU using a random variable z_2 that is correlated with the test results Y so that the Spearman correlation between the sampling weights, and the true test results is -0.06 (the details are described in Section 3 of the Supplementary Material). For both testing a single trait and 2 traits and for both noninformative and informative sampling weights, the coefficient of variation of the resulting sampling weights ranges from 60.5% to 69.2%. The 5 different methods M1 to M5 for grouping the data with a fixed group size of 5 are called random grouping, systematic grouping, serpentine grouping, two-strata grouping, and three-strata grouping as described in section 2.2.

To appreciate the need to incorporate the sampling weights in the estimation, prevalence estimates are reported in Table 2 when the sampling weights are *ignored* even though they are informative and testing is done for each individual as well as in groups that were randomly formed (M1). Notice the substantial bias in the prevalence estimates especially relative to the standard errors. These results demonstrate that the sampling weights under informative sampling need to be incorporated into the estimator of prevalence. Under informative sampling, the mean squared errors are greater when $\pi = 7\%$ than $\pi = 1\%$, and it results in very low coverage probabilities. In this case for grouped data, the bias means that the expected value of the estimate of prevalence is 8.67% rather than 7.0% and the average half length of the 95% confidence interval is 1.18. This implies that most of the confidence intervals will not include 7%. Because the magnitude of the bias relative to the standard errors is much less when the prevalence is 1%, the coverage probability is closer to the nominal value. Similar results occur for the coverage probabilities for informative weights for the ungrouped case in Table 2. When sampling weights are noninformative, the estimates in groups are slightly biased.

The second simulation scenario shows that incorporating the sampling weights in the estimations for a single trait and 2 traits does not influence the bias of the estimator but slightly increases its variance when the weights are noninformative (Tables 2 and 3). There are essentially no differences in bias and variance among the 5 grouping methods (Tables 3 and 4) as expected. Hence, all of the 5 grouping methods with the average weights have good performance in terms of the point and variance estimation. In contrast, when sampling weights are informative, the 2-strata/3-strata grouping methods (Table 3) and the serpentine method (Table 5) outperformed the other methods for the prevalence estimations of a trait and 2 traits, respectively. The prevalence estimates provided by grouping methods M4/M5 for a single trait and M3 for 2 traits are nearly unbiased. The best grouping method depends on the degree of variation in the sample weights.

When compared with the standard errors from ungrouped data, there is greater efficiency (ie, smaller standard errors) for estimating a single prevalence and the slightly lower efficiency for simultaneously estimating 2 prevalences. The reason that the estimates for π_{00} are relatively more biased than those for $\pi_{10}(\pi_{01})$ and π_{11} is because the simulation used a higher sampling fraction for responses with a single positive test or with 2 positive tests.

The third simulation scenario shows that we need to be cautious about referring to external information for the extent of misclassification of a test. We conceive of 3 cases: (1) We

know only confidence intervals for sensitivity and specificity, say [70–90] for 80% sensitivity and [90–100] for 95% specificity; (2) we know only the values of sensitivity or specificity but not both; (3) we assume the test has no error. We use the proposed methods where sensitivity and/or specificity values are misspecified, and the results are summarized in Tables S2 to S5. When the sampling weights are noninformative, we focus on the M1 grouping. We found the impact of misspecifying sensitivity and specificity values on the estimation from grouped data is less than the ones from ungrouped data. Even when the difference from the true specificity is 5%, the bias is substantial, so the estimator is sensitive to the sensitivity and specificity values.

In summary, when the assumed sensitivity and specificity are correct, the 5 grouping methods are equivalent for noninformative sampling weights, whereas for the informative sampling weights, the M4/M5 for estimating a single trait and M3 for estimating 2 traits perform the best.

4 | APPLICATION

We applied the proposed grouping methods to the data from the NHANES in 2009 to 2012 for estimating the prevalence of oral HPV infection in the United States (US) population. The design for NHANES 2009 to 2012 was a stratified multistage probability sample of the civilian, noninstitutionalized US population. Oral HPV infection is newly appreciated as a strong risk factor for a distinct type of oropharyngeal squamous cell carcinoma that has increasing incidence in the US.^{24,25} Oral rinse specimens collected from men and women 14 to 69 years were processed, stored, and shipped to the Ohio State University Gillison Laboratory, Columbus, Ohio.²⁶ The oral HPV rinse data are from 62 sampled PSUs across 29 sampling strata, and 2 PSUs are selected from the strata except for 3 strata, where 3 PSUs were sampled. We analyzed data from 9256 individuals with complete assayed oral HPV DNA types. The sample is well-balanced with respect to sex, with 4600 males vs 4656 females (weighting counts = 185,854,268 49.9%, vs 186,951,423 50.1%). Thirty-seven HPV subtypes were detected in the sample by multiplex polymerase chain reaction (PCR).²⁴ Samples were reported as HPV negative if none of the 37 HPV DNA subtypes are positive. In this analysis, we focus on the estimation of the 14 oncogenetic HPV DNA subtypes (HPV 16, 18, 45, 31, 33, 58, 52, 51, 39, 68, 35, 59, 56, and 66). Among those subtypes, HPV16 and HPV18 are known to be cancer-causing/high-risk subtypes, so we also consider jointly estimating the prevalence of the 2 subtypes.

Since the Spearman correlations between the sampling weights and responses, oncogenetic HPV (a binary trait) and HPV16 and/or HPV18 infections (4 categorical multinomial trait) are small $\rho_{onco} = -0.001$, $\rho_{16\&18} = 0.009$, indicating that the sample weights are not very informative. However, for comparison, we use both the weighted and unweighted-generalized estimating equation for these responses.

The NHANES data includes individual test results for the 37 HPV DNA subtypes; therefore, we can compare the prevalence estimates for grouped data with the ones for ungrouped data. The sensitivity and specificity for the bioassay test for ungrouped and grouped oral HPV rinse are not reported. We use the external sensitivity and specificity values of 99.5% and

100%, respectively²⁷ and assume that the misclassification values for the grouped data are identical to those for the ungrouped data. For contrast, we also estimate the prevalences assuming the test to be 100% accurate. The estimated prevalence of oncogenetic oral HPV infection for the grouped data is about 4.05% (SE = 0.26) and the prevalences of HPV 16/18 only and both, 16 and 18 are estimated to be 0.94% (SE = 0.10), 0.25% (SE = 0.05), and 0.02% (SE = 0.02), respectively, under correct specified sensitivity and specificity. The estimated prevalences for the grouped data are similar to the ones for the ungrouped data (Table 6), demonstrating the potential usefulness of grouping survey data when assuring confidentiality is important.

5 | DISCUSSION

In this paper, we consider the problem of how to use group testing of assays collected from individuals sampled from complex sample surveys while protecting the confidentiality of respondents. Confidentiality of respondents is maintained because individuals are tested in groups and no individual is ever tested. If the sampling weights are noninformative, we can ignore the weights in analyzing data; however, in practice, it is difficult to judge whether the sampling weights are informative (Korn and Graubard¹⁸, pp196–199). Hence, we need a group test weighted estimator, and to recapture some of the information from the sampling weights, we propose using the average weights of the observations in each group as the weight in the estimation. Grouping data based on sampling weights when the sampling weights are informative has similarities to the covariates-homogeneous grouping approach of Mitchell et al.²⁸

We adopted average-weighted generalized estimating equations using an independent working correlation for a binary trait or 2 binary traits tested in groups and account for the intraclass correlation of the cluster sample design in the variance estimation. We consider the 5 different methods of grouping observations within each PSU. The simulations indicated that the accuracy of the proposed estimators using these grouping methods depends on the degree of informativeness of the sample weights and the magnitude of intercluster correlation of the prevalence. When the sample weighting is noninformative, the nonrandom grouping because of nesting the grouping within PSUs resulted in slightly biased estimation of prevalences.

The grouping methods were empirically evaluated by simulations and a data analysis of the prevalence of oral HPV positivity in the NHANES 2009 to 20012. The sampling weights in the NHANES data seem to be not very informative with respect to the oral HPV infection. None of the 5 grouping methods is uniformly the best in terms of mean square error when the sample weights are informative. When designing a group testing procedure, we do not know the informativeness of the sampling weights because traits have not yet been observed; however, if we know of covariates that are strongly associated with the traits, we can calculate the correlations between the covariates and sampling weights to obtain some information about the possible informativeness of sample weights. This should assist researchers in their choice of an appropriate group testing method.

Three features of the group testing need to be considered in order to determine an appropriate group size for the prevalence of rare traits: (a) dilution effect that can result from the combining of bio-specimens in each group, eg, a single positive sample would be diluted if the rest of the group is negative; (b) the effect of combining bio-specimens on the testing of the grouped sample, eg, unpredictable biological consequences may result from mixing samples of different people; and (c) group size should be large enough to protect confidentiality; features (a) and (b) can result in misclassification for the group testing. Tradeoffs between the nature and the sensitivity of the trait under study, the required degree (level) of accuracy of the estimate and the degree of confidentiality of the respondents that were promised, should be considered when determining the group size.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

1. Dorfman R The detection of defective members of large populations. *Ann Math Stat.* 1943;14(4): 436–440.
2. Finucan HM. The blood testing problem. *J R Stat Soc Ser C.* 1964;13(1):43–50.
3. Litvak E, Tu XM, Pagano M. Screening for the presence of a disease by pooling sera samples. *J Am Stat Assoc.* 1994;89:424–434.
4. Sobel M, Elashoff RM. Group testing with a new goal, estimation. *Biometrika.* 1975;62:181–193.
5. Barillot E, Lacroix B, Cohen D. Theoretical analysis of library screening using a n-dimensional pooling strategy. *Nucleic Acids Res.* 1991;19:6241–6247. [PubMed: 1956784]
6. Phatarford RM, Sudbury A. The use of a square array scheme in blood testing. *Stat Med.* 1994;13:2337–2343. [PubMed: 7855467]
7. Kim HY, Hudgens MG. Three-dimensional array-based group testing algorithms. *Biometrics.* 2009;65:903–910. [PubMed: 19053996]
8. Hwang FK. Group testing with a dilution effect. *Biometrika.* 1976;63:671–683.
9. McMahan CS, Tebbs JM, Bilder CR. Regression models for group testing data with pool dilution effects. *Biostatistics.* 2013;14:284–298. [PubMed: 23197382]
10. Liu A, Liu C, Zhang Z, Albert PS. Optimality of group testing in the presence of misclassification. *Biometrika.* 2012;99:245–251. [PubMed: 23049137]
11. Cahoon-Young B, Chandler A, Livermore T, Gaudino J, Benjamin R. Sensitivity and specificity of pooled versus individual sera in a human immunodeficiency virus antibody prevalence study. *J Clin Microbiol.* 1989;27:1893–1895. [PubMed: 2504779]
12. Gastwirth JL, Hammick PA. Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of aids antibodies in blood donors. *J Stat Plan Inference.* 1989;22:15–27.
13. Hammick PA, Gastwirth JL. Group testing for sensitive characteristics: extension to higher prevalence levels. *Int Stat Rev; Internat Statist Rev.* 1994;62(3):319–331.
14. Pfeiffer RM, Rutter JL, Gail MH, Struwing J, Gastwirth JL. Efficiency of dna pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genet Epidemiol.* 2002;22:94–102. [PubMed: 11754476]
15. National Center for Health Statistics (NCHS) consent forms and confidentiality brochure. Available at https://www.ignorespacescdc.gov/nchs/data/nhanes/2009-2010/documents/english_confidentiality.pdf. Accessed December 10, 2017.
16. Caudill SP. Use of pooled samples from the national health and nutrition examination survey. *Stat Med.* 2012;31(27):3269–3277. [PubMed: 22492247]

17. Caudill SP. Confidence interval estimation for pooled-sample biomonitoring from a complex survey design. *Environ Int.* 2015;85:40–45. [PubMed: 26313704]
18. Korn EL, Graubard BI. *Analysis of Health Survey*. Wiley; 1989.
19. Graubard BI, Fears TR. Standard errors for attributable risk for simple and complex sample designs. *Biometrics.* 2005;61:847–55. [PubMed: 16135037]
20. Nash JC, Varadhan R. Unifying optimization algorithms to aid software system users: optimx for R. *J Stat Softw.* 2011;43(9):1–14.
21. Curtin LR, Mohadjer LK, Dohrmann SM, et al. The national health and nutrition examination survey: sample design, 2007–2010. *Vital Health Stat.* 2013;2:1–23.
22. Johnson CL, Dohrmann SM, Burt VL, Mohadjer LK. National health and nutrition examination survey: sample design, 2011–2014. *Vital Health Stat.* 2014;2:1–33.
23. Brier S. Analysis of contingency tables under cluster sampling. *Biometrika.* 1980;67(3):591–596.
24. Gillison ML, Broutian T, Pickard RKL, et al. Prevalence of oral HPV infection in the United States, 2009–2010. *J Am Med Assoc.* 2012;307(7):693–703.
25. Chaturvedi AK, Katki HA, Hildesheim A, et al. Human papillomavirus infection with multiple types: pattern of coinfection and risk of cervical disease. *J Infect Dis.* 2011;203(7):910–920. [PubMed: 21402543]
26. National Health and Nutrition Examination Survey. Human papillomavirus (hpv) – oral rinse (ORHPV_F). https://www.cdc.gov/nchs/nhanes/2009-2010/ORHPV_F.htm. Accessed December 10, 2017.
27. Walline KC, McHugh JB, Byrd SA, et al. High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and oral cavity cancers: comparison of multiple methods. *JAMA Otolaryngol Head Neck Surg.* 2013;139:1320–1327. [PubMed: 24177760]
28. Mitchell EM, Lyles RH, Manatunga AK, Danaher M, Perkins NJ, Schisterman EF. Regression for skewed biomarker outcomes subject to pooling. *Biometrics.* 2014;70:202–211. [PubMed: 24521420]

FIGURE 1.

Each number in the circles represents a measure of size. PSU, primary sampling unit

FIGURE 2.

The numbers in the circles represent the sampling weights; gp stands for group; Strata A, B, and C are different from sampling strata $l = 1, \dots, L$; the 5 grouping methods are done within each cluster; M1 = randomly group observations; M2 = sort observations by their sample weights and group q adjacent observations; M3 = group the sorted observations by their sampling weights in a serpentine fashion; M4 = first stratify the observations into high/low sampling weight strata and then randomly form groups of observations within strata; M5 = the same as M4 but stratify into 3 categories, low/medium/high

TABLE 1

Simulation scenarios

Scenario	Inference procedure	Sampling weight	Response	Grouping method	Table number
1	Ignoring weights with known sensitivity and specificity values	Not informative	Binomial	M1	Table 2
		Informative	Binomial	M1	Table 2
2	Weighted GEE with known sensitivity and specificity values	Not informative	Binomial	M1 to M5	Table 3
			Multinomial	M1 to M5	Table 4
		Informative	Binomial	M1 to M5	Table 3
			Multinomial	M1 to M5	Table 5
3	Weighted GEE and incorrectly specifying sensitivity and specificity values	Not informative	Binomial	M1 to M5	Table S2
			Multinomial	M1 to M5	Table S3
		Informative	Binomial	M1 to M5	Table S4
			Multinomial	M1 to M5	Table S5

Abbreviation: GEE, generalized estimating equation.

TABLE 2

Simulation results when the sampling weights are *ignored*, and true sensitivity and specificity are 80% and 95%, respectively^a

Label	<u>Not informative Weight</u>					<u>Informative Weight</u>				
	Bias, %	SE	ASE	MSE	CP, 95%	Bias, %	SE	ASE	MSE	CP, 95%
True prevalence = 7.0%										
Ungrouped data	-0.01	0.53	0.53	0.29	94.3	1.91	0.62	0.62	4.04	14.6
Grouped data (M1)	-0.13	0.48	0.48	0.25	92.8	1.67	0.59	0.58	3.14	18.6
True prevalence = 1.0%										
Ungrouped data	-0.01	0.35	0.35	0.13	93.5	0.3	0.38	0.37	0.24	89.0
Grouped data (M1)	-0.02	0.22	0.22	0.05	93.0	0.27	0.26	0.25	0.15	85.5

Abbreviations: ASE, asymptotic standard error; CP, coverage probability; MSE, mean squared error; SE, standard error.

^aRandomly grouping method is used, and the group size = 5.

TABLE 3

Simulation results of the weighted GEE method for binary responses when the true sensitivity and specificity are 80% and 95%, respectively^a

Label	Not informative Weight					Informative Weight				
	Bias, %	SE	ASE	MSE	CP, 95%	Bias, %	SE	ASE	MSE	CP, 95%
True prevalence = 7.0%										
Ungrouped data	0.00	0.61	0.61	0.38	94.2	0.01	0.60	0.59	0.36	95.1
Grouped data										
(M1) randomly gp	-0.13	0.52	0.51	0.29	92.1	0.86	0.56	0.55	1.05	69.9
(M2) systematic gp	-0.14	0.54	0.55	0.32	92.6	-0.34	0.51	0.51	0.38	89.0
(M3) serpentine gp	-0.13	0.51	0.51	0.28	92.8	1.12	0.57	0.56	1.58	53.2
(M4) 2-strata gp	-0.12	0.53	0.53	0.30	93.5	0.07	0.53	0.53	0.29	95.2
(M5) 3-strata gp	-0.13	0.54	0.53	0.31	92.7	-0.11	0.53	0.52	0.29	93.8
True prevalence = 1.0%										
Ungrouped data	0.00	0.42	0.41	0.18	94.0	0.00	0.40	0.39	0.16	95.1
Grouped data										
(M1) randomly gp	-0.02	0.24	0.23	0.06	92.3	0.14	0.24	0.24	0.08	93.3
(M2) systematic gp	-0.02	0.25	0.25	0.07	93.3	-0.03	0.24	0.23	0.06	93.2
(M3) serpentine gp	-0.02	0.23	0.23	0.06	93.1	0.17	0.25	0.24	0.09	91.4
(M4) 2-strata gp	-0.02	0.24	0.24	0.06	92.9	0.03	0.23	0.23	0.05	95.2
(M5) 3-strata gp	-0.02	0.24	0.24	0.06	93.1	0.01	0.24	0.23	0.06	95.1

Abbreviations: ASE, asymptotic standard error; CP, coverage probability; GEE, generalized estimating equation; MSE, mean squared error; SE, standard error.

^aThe group size = 5.

TABLE 4

Simulation results of the weighted GEE method for multinomial responses when the true sensitivity and specificity are 80% and 95%, respectively, and the sampling weights are not informative^a

Label	True Value, %	Bias, %	SE	ASE	MSE	CP, 95%
Ungrouped data						
	$\bar{\pi}_{00} = 92.36$	0.01	0.74	0.73	0.55	94.5
	$\bar{\pi}_{10} = 3.25$	0.00	0.52	0.52	0.27	94.7
	$\bar{\pi}_{01} = 3.25$	0.00	0.53	0.52	0.28	94.7
	$\bar{\pi}_{11} = 1.14$	0.00	0.27	0.27	0.07	94.2
Grouped data						
(M1) randomly grouping	$\bar{\pi}_{00} = 92.36$	0.14	0.56	0.56	0.33	93.5
	$\bar{\pi}_{10} = 3.25$	-0.06	0.41	0.40	0.17	93.9
	$\bar{\pi}_{01} = 3.25$	-0.06	0.40	0.40	0.16	94.1
	$\bar{\pi}_{11} = 1.14$	-0.02	0.27	0.26	0.07	94.0
(M2) systematic grouping	$\bar{\pi}_{00} = 92.36$	0.14	0.60	0.60	0.38	94.0
	$\bar{\pi}_{10} = 3.25$	-0.05	0.44	0.44	0.20	94.0
	$\bar{\pi}_{01} = 3.25$	-0.06	0.44	0.44	0.20	94.0
	$\bar{\pi}_{11} = 1.14$	-0.03	0.29	0.29	0.08	93.7
(M3) serpentine grouping	$\bar{\pi}_{00} = 92.36$	0.14	0.55	0.55	0.32	93.9
	$\bar{\pi}_{10} = 3.25$	-0.06	0.40	0.40	0.16	93.6
	$\bar{\pi}_{01} = 3.25$	-0.06	0.40	0.40	0.16	94.1
	$\bar{\pi}_{11} = 1.14$	-0.02	0.27	0.26	0.07	93.8
(M4) 2-strata grouping	$\bar{\pi}_{00} = 92.36$	0.14	0.57	0.58	0.34	93.6
	$\bar{\pi}_{10} = 3.25$	-0.06	0.42	0.42	0.18	94.5
	$\bar{\pi}_{01} = 3.25$	-0.06	0.42	0.42	0.18	94.4
	$\bar{\pi}_{11} = 1.14$	-0.02	0.28	0.27	0.08	94.2
(M5) 3-strata grouping	$\bar{\pi}_{00} = 92.36$	0.13	0.59	0.58	0.36	94.0
	$\bar{\pi}_{10} = 3.25$	-0.05	0.43	0.42	0.19	94.2
	$\bar{\pi}_{01} = 3.25$	-0.06	0.42	0.43	0.18	94.1
	$\bar{\pi}_{11} = 1.14$	-0.02	0.28	0.28	0.08	94.1

Abbreviations: ASE, asymptotic standard error; CP, coverage probability; GEE, generalized estimating equation; MSE, mean squared error; SE, standard error.

^aThe group size = 5.

TABLE 5

Simulation results of the weighted GEE method for multinomial responses when the true sensitivity and specificity are 80% and 95%, respectively, and the sampling weights are informative^a

Label	True Value, %	Bias, %	SE	ASE	MSE	CP, 95%
Ungrouped data						
	$\bar{\pi}_{00} = 92.36$	-0.01	0.71	0.70	0.50	95.3
	$\bar{\pi}_{10} = 3.25$	0.00	0.49	0.49	0.24	95.1
	$\bar{\pi}_{01} = 3.25$	0.00	0.50	0.50	0.25	95.0
	$\bar{\pi}_{11} = 1.14$	0.01	0.25	0.25	0.06	94.9
Grouped data						
(M1) randomly grouping	$\bar{\pi}_{00} = 92.36$	-0.26	0.57	0.56	0.39	93.9
	$\bar{\pi}_{10} = 3.25$	0.09	0.41	0.41	0.18	94.8
	$\bar{\pi}_{01} = 3.25$	0.10	0.41	0.41	0.18	94.7
	$\bar{\pi}_{11} = 1.14$	0.07	0.27	0.27	0.08	95.4
(M2) systematic grouping	$\bar{\pi}_{00} = 92.36$	3.14	0.37	0.36	10.00	0.0
	$\bar{\pi}_{10} = 3.25$	-1.84	0.30	0.29	3.48	0.0
	$\bar{\pi}_{01} = 3.25$	-1.90	0.29	0.29	3.69	0.0
	$\bar{\pi}_{11} = 1.14$	0.60	0.25	0.25	0.42	34.3
(M3) serpentine grouping	$\bar{\pi}_{00} = 92.36$	-0.16	0.58	0.58	0.36	94.9
	$\bar{\pi}_{10} = 3.25$	0.07	0.41	0.41	0.17	95.0
	$\bar{\pi}_{01} = 3.25$	0.06	0.42	0.41	0.18	95.0
	$\bar{\pi}_{11} = 1.14$	0.03	0.27	0.27	0.07	94.7
(M4) two-strata grouping	$\bar{\pi}_{00} = 92.36$	-0.39	0.60	0.58	0.51	90.7
	$\bar{\pi}_{10} = 3.25$	0.13	0.43	0.42	0.20	94.9
	$\bar{\pi}_{01} = 3.25$	0.14	0.43	0.43	0.20	94.4
	$\bar{\pi}_{11} = 1.14$	0.12	0.28	0.28	0.09	94.6
(M5) three-strata grouping	$\bar{\pi}_{00} = 92.36$	-0.35	0.60	0.59	0.48	92.0
	$\bar{\pi}_{10} = 3.25$	0.08	0.43	0.43	0.19	95.0
	$\bar{\pi}_{01} = 3.25$	0.09	0.43	0.43	0.19	95.2
	$\bar{\pi}_{11} = 1.14$	0.17	0.29	0.29	0.11	93.5

Abbreviations: ASE, asymptotic standard error; CP, coverage probability; GEE, generalized estimating equation; MSE, mean squared error; SE, standard error.

^aThe group size = 5.

TABLE 6

NHANES 2009 to 2012^a

Label	Oncogenetic HPV		HPV 16		HPV 18		HPV 16 & 18	
	Est	SE	Est	SE	Est	SE	Est	SE
Ungrouped data, sensitivity = 100%, specificity = 100%								
Weighted	4.08	0.34	1.00	0.12	0.24	0.06	0.06	0.04
Not weighted	4.03	0.26	0.94	0.10	0.25	0.05	0.02	0.02
Grouped data, sensitivity = 100%, specificity = 100%								
(M1) random grouping	4.13	0.18	1.10	0.08	0.28	0.04	0.03	0.02
(M2) systematic grouping	4.12	0.23	1.06	0.08	0.21	0.05	0.07	0.04
(M3) serpentine grouping	4.11	0.19	0.99	0.07	0.27	0.04	0.03	0.02
(M4) 2-strata gp	4.08	0.25	1.01	0.08	0.24	0.04	0.04	0.03
(M5) 3-strata gp	4.21	0.26	1.04	0.09	0.21	0.04	0.06	0.03
Ungrouped data, sensitivity = 99.5%, specificity = 100%								
Weighted	4.10	0.34	1.01	0.13	0.24	0.06	0.06	0.04
Not weighted	4.05	0.26	0.94	0.10	0.25	0.05	0.02	0.02
Grouped data, sensitivity = 99.5%, specificity = 100%								
(M1) random grouping	4.30	0.23	1.06	0.08	0.30	0.05	0.03	0.03
(M2) systematic grouping	4.14	0.23	1.06	0.08	0.21	0.05	0.07	0.04
(M3) serpentine grouping	4.13	0.19	0.99	0.07	0.27	0.04	0.03	0.02
(M4) 2-strata gp	4.18	0.23	1.01	0.08	0.24	0.04	0.04	0.02
(M5) 3-strata gp	4.26	0.28	1.05	0.09	0.22	0.04	0.05	0.03

Abbreviations: Est, estimate; HPV, human papillomavirus; SE, standard error.

^aOral HPV infection by DNA type.