Hindawi
Computational and Mathematical Methods in Medicine
Volume 2019, Article ID 4381084, 8 pages
https://doi.org/10.1155/2019/4381084

Research Article

Determination of Varying Group Sizes for Pooling Procedure

Wenjun Xiong, Hongyu Lu, and Juan Ding

School of Mathematics and Statistics, Guangxi Normal University, Yucai Road 15, Guilin 541004, China

Correspondence should be addressed to Juan Ding; dingjuan@gxnu.edu.cn

Received 15 May 2018; Revised 17 January 2019; Accepted 5 February 2019; Published 1 April 2019

Academic Editor: Nadia A. Chuzhanova

Pooling is an attractive strategy in screening infected specimens, especially for rare diseases. An essential step of performing the pooled test is to determine the group size. Sometimes, equal group size is not appropriate due to population heterogeneity. In this case, varying group sizes are preferred and could be determined while individual information is available. In this study, we propose a sequential procedure to determine varying group sizes through fully utilizing available information. This procedure is data driven. Simulations show that it has good performance in estimating parameters.

1. Introduction

Routine monitoring or large scale of screening usually occurs in biomedical research to identify infected specimens [1–4]. However, some test kits, e.g., nucleic acid amplification test (NAAT), are expensive [2, 5]. Therefore, the expense during a large-scale monitoring process is usually a financial burden if resource is limited [6–8]. The strategy of pooling biospecimens is attractive to address this issue [9–11], which was first used during World War II to screen for syphilis [12]. This strategy is firstly to pool specimens into groups and then screen these groups. If a group tests negative, all specimens in this group will be declared negative; otherwise, continue to perform individual test. When the prevalence is low, the total number of tests using pooling will be far less than that using the individual test. Due to its efficiency and cost saving, pooling is now applied in many fields, such as agriculture [13], genetics [14, 15], HIV/AIDS [16, 17] and blood screening [18], and environmental epidemiology [19, 20].

The gain of pooling mainly depends on the pooling algorithm. Assuming homogeneity of the population, dozens of papers have investigated the problem how to design an efficient algorithm [21–25]. However, this assumption might be violated in practical application [26–28]. While individual information is available, it is of interest to estimate individual-level prevalence through incorporating such information. Note that only group-level status is observed, e.g., positive or negative. This problem has been studied in parametric context through the framework of binary regression models [29–31], and also in semiparametric [32, 33] or nonparametric context [34, 35]. However, aforementioned work mostly uses a single group size that is determined in advance.

A set of pool sizes might be more appropriate while considering population heterogeneity. For example, varying pool sizes were used to estimate the infection prevalence of *Myxobolus cerebralis*, which causes whirling disease, among free-ranging salmonid fish collected from the Truckee River in Nevada and California [36]. In a study of estimating the prevalence of several viruses in carnations grown in nursery glasshouses in Victoria, sequential pooled testing involving several pool sizes was adopted [37]. Using a single group size might be optimal for some estimates but far from others, especially when we have little information ahead of the experiment [37, 38]. More work is better on this issue since the benefit of pooling algorithm mainly depend on the choice of pool size [38–40]. In this study, we propose a pooling strategy with varying pool sizes through taking advantage of individual information. Our procedure is a data-driven pooling algorithm, where groups are formed sequentially. Its performance is extensively investigated by simulations and a real data set.

Table 1: The performance of estimators using different pooling procedures.

(S_e, S_p)	$\mathscr{A}$	$m = 1000$				$m = 500$			
		β_0		β_1		β_0		β_1	
		Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
X ~ N (2, 1.5)									
(0.99, 0.99)	PSV	−3.003	0.020	0.401	0.002	−3.001	0.043	0.401	0.004
	PSF(k^*)	−3.002	0.010	0.402	0.002	−3.006	0.022	0.403	0.004
	PSF(5)	−3.007	0.134	0.402	0.010	−3.018	0.289	0.405	0.021
	PSF(10)	−3.006	0.021	0.403	0.003	−3.009	0.042	0.403	0.005
(0.95, 0.95)	PSV	−3.002	0.026	0.402	0.003	−2.999	0.050	0.401	0.005
	PSF(k^*)	−3.006	0.022	0.403	0.003	−3.009	0.041	0.406	0.006
	PSF(5)	−3.008	0.162	0.403	0.012	−2.997	0.317	0.400	0.023
	PSF(10)	−3.004	0.026	0.403	0.003	−2.998	0.052	0.401	0.006
(0.9, 0.9)	PSV	−3.001	0.034	0.402	0.003	−2.991	0.071	0.395	0.007
	PSF(k^*)	−3.004	0.035	0.404	0.004	−3.007	0.074	0.404	0.010
	PSF(5)	−2.974	0.225	0.394	0.016	−2.993	0.418	0.399	0.031
	PSF(10)	−3.004	0.038	0.404	0.005	−3.008	0.077	0.404	0.010
X ~ Γ (2.5, 0.8)									
(0.99, 0.99)	PSV	−2.991	0.041	0.397	0.004	−2.997	0.020	0.399	0.002
	PSF(k^*)	−3.006	0.020	0.404	0.003	−3.002	0.010	0.402	0.002
	PSF(5)	−2.973	0.281	0.393	0.020	−3.002	0.136	0.400	0.010
	PSF(10)	−3.002	0.042	0.402	0.005	−3.004	0.021	0.402	0.002
(0.95, 0.95)	PSV	−3.000	0.053	0.401	0.005	−2.998	0.026	0.400	0.003
	PSF(k^*)	−3.010	0.041	0.404	0.006	−3.007	0.020	0.404	0.003
	PSF(5)	−3.060	0.324	0.416	0.023	−3.015	0.171	0.405	0.012
	PSF(10)	−3.003	0.053	0.402	0.007	−3.006	0.027	0.403	0.003
(0.9, 0.9)	PSV	−2.989	0.072	0.398	0.007	−2.992	0.034	0.399	0.004
	PSF(k^*)	−3.017	0.075	0.408	0.010	−3.001	0.033	0.402	0.004
	PSF(5)	−3.012	0.379	0.403	0.028	−2.995	0.198	0.398	0.014
	PSF(10)	−3.018	0.075	0.409	0.010	−3.003	0.035	0.402	0.005

2. Methods

2.1. Notations and Background. Suppose N specimens are assigned into m groups each with size k_i for $i = 1, 2, \ldots, m$. z_i denotes the observed status of the i^{th} group, and X_{ij} denotes the covariates of the j^{th} specimen in the i^{th} group for $j = 1, \ldots, k_i$ and $i = 1, \ldots, m$. The observations are $\{z_i, X_{ij}, j = 1, \ldots, k_i, i = 1, \ldots, m\}$, where $X_{ij} = \{1, x_{1,ij}, \ldots, x_{d-1,ij}\}^{\mathrm{T}}$. Here, the notation A^{T} represents the transpose of matrix A. The sensitivity and specificity of the screening tool are denoted by S_e and S_p, respectively. The full likelihood function is

$$L(\beta; z, X) = \prod_{i=1}^{m} \left[S_e - r \prod_{j=1}^{k_i} \left(1 - p_{ij}\right) \right]^{z_i} \cdot \left[1 - S_e + r \prod_{j=1}^{k_i} \left(1 - p_{ij}\right) \right]^{1-z_i}, \tag{1}$$

where $r = S_e + S_p - 1$ and $p_{ij} = g(\beta_0 + \beta_1 x_{1,ij} + \cdots + \beta_{d-1} x_{d-1}, ij) = g(X_{ij}^{\mathrm{T}}\beta)$. The parameter β is defined by $\beta = \{\beta_0, \beta_1, \ldots, \beta_{d-1}\}^{\mathrm{T}}$, and the function $g^{-1}(\cdot)$ is a known, monotone, and differentiable link function.

Sometimes there might be a maximum admissible group size $k^{\max}$, e.g., a large group size might bring the dilution effect. Therefore, we should carefully choose an appropriate group size that is smaller than $k^{\max}$. Define a set $\mathscr{K} = \{1, 2, \ldots, k^{\max}\}$, and denote it by $\mathbf{k} = \{k_1, \ldots, k_m\}$, $k_i \in \mathscr{K}$, $i = 1, \ldots, m$. Once the group size $\mathbf{k}$ is determined, we could obtain the estimator of β through maximum likelihood function $L(\beta, z, X)$. The Fisher information matrix of the parameter β could be rewritten as follows:

$$I(\beta, \mathbf{k}) = \sum_{i=1}^{m} \frac{G_i(k_i, \beta) G_i^{\mathrm{T}}(k_i, \beta)}{C_i(\beta, k_i)}, \tag{2}$$

where

$$H_i(k_i, \beta) = -\frac{1}{k_i} \sum_{j=1}^{k_i} \log\left(1 - g\left(X_{ij}^{\mathrm{T}}\beta\right)\right),$$

$$G_i(k_i, \beta) = \frac{\partial}{\partial \beta} H_i(k_i, \beta),$$

$$C_i(\beta, k_i) = \left(S_e - r\exp^{-k_i H_i(k_i, \beta)}\right)\left(1 - S_e + r\exp^{-k_i H_i(k_i, \beta)}\right) \cdot r^{-2} k_i^{-2} \exp^{2k_i H_i(k_i, \beta)}. \tag{3}$$

The calculation of Fisher information $I(\beta, \mathbf{k})$ is presented in Supplemental Material (Available here). To obtain a better estimator $\widehat{\beta}$, we try to find $\mathbf{k}$ that maximizes Fisher information $I(\beta, \mathbf{k})$. However, individual-level measurements make it difficult to achieve this goal.

The Fisher information $I(\beta, \mathbf{k})$ defined in (2) involves a measurement $H_i(\beta, k_i)$, along with its functions $G_i(k_i, \beta)$

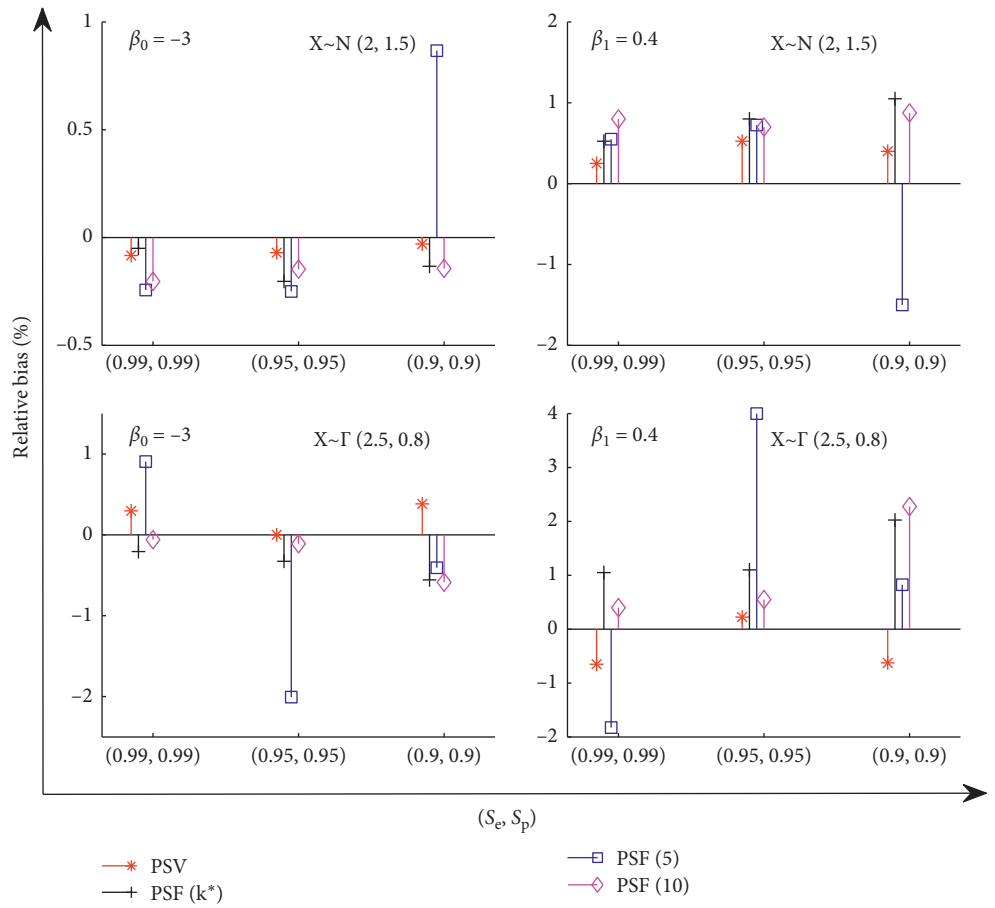

FIGURE 1: The relative bias of the parameters β_0 and β_1. The distribution of covariant is set by $N(2, 1.5)$ (top two panels) and $\Gamma(2.5, 0.8)$ (bottom two panels), with the fixed number of groups $m = 1000$.

and $C_i(\beta, k_i)$. According to Delaigle and Hall [41], $\prod_{j=1}^{k_i}(1-g(X_{ij}^{\mathrm{T}}\beta))$ is generally close to $(1-g(\overline{X}_i^{\mathrm{T}}\beta))^{k_i}$, where $\overline{X}_i = 1/k_i\sum_{j=1}^{k_i}X_{ij}$. This closeness let the Fisher information reduce to the following format: $I(\beta, \mathbf{k}) = \sum_{i=1}^{m}Z_i(\beta)Z_i(\beta)^{\mathrm{T}}/C_i(\beta, k_i)$, where $Z_i(\beta) = g'(\overline{X}_i^{\mathrm{T}}\beta)\overline{X}_i/(1-g(\overline{X}_i^{T}\beta))$. Then, we propose to determine the group sizes through minimizing all $C_i(\beta, k_i)$ with respect to k_i for $i = 1, \ldots, m$.

Note that the aforementioned approximate approach requires the pools are homogeneous. There are two methods to obtain homogeneous pool: reorder the specimens according to similarity of covariants or based on individual risk probability. The latter is adopted in this study. Following the method in McMahan et al. [42], the procedure of forming homogeneous pool is as follows. Firstly, use training data or prior knowledge to obtain an initial estimator $\beta^{(0)}$ [42]. Secondly, sort the specimens by their risk probability. Let G denotes the set which contains total covariants of enrolled specimens, $G = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where N is the number of specimens and $\mathbf{x}_i$ is the covariant of the i^{th} specimen. Sort G by risk probability $p_i = g(\mathbf{x}_i^{\mathrm{T}}\beta^{(0)})$ in the descending order, and obtain a sorted set $G^{\mathrm{s}} = \{\mathbf{x}_1^{\mathrm{s}}, \cdots, \mathbf{x}_N^{\mathrm{s}}\}$. The remaining procedure is directly performed on this sorted set.

2.2. Sequential Adaptive Pooling Algorithm. Our strategy is an adaptive design, which is often adopted in the biological experiment and also in the pooled test [22]. Before stating the algorithm, we need the following result. Suppose the specimens are assigned for the first $l-1$ groups with the corresponding group sizes $\{k_1, \ldots, k_{l-1}\}$. Let $n_l = \sum_{j=1}^{l}k_j$ for $l \geq 1$ and $n_0 = 0$. Denote $W_l(\beta) = -\log(1-g((\mathbf{x}_{n_{l-1}+1}^{\mathrm{s}})^{\mathrm{T}}\beta))$. Then the group size for the next group, k_l, equals $k^{\max}$ if $k^{\max} \leq \phi_0/W_l(\beta^{(0)})$. Here, ϕ_0 is the root of an equation $2S_{\mathrm{e}}(1-S_{\mathrm{e}})(\phi-1)e^{2\phi} + r(2S_{\mathrm{e}}-1)(\phi-2)e^{\phi} + 2r^2 = 0$ and is approximately 1.8414. The proof of this result is presented in Supplemental Material (Available here). Our pooling strategy is described as follows:

Step 1. Label the specimens according to the ordering of G^{s}. For example, label the specimen with covariants $\mathbf{x}_1^{\mathrm{s}}$ by number 1. Assign specimens with labels up to $k^{\max}$ into l^{th} group.

Step 2. Calculate the corresponding function $C_l(\beta^{(0)}, k), k \in \mathscr{K}$ and $c_0 = \phi_0/W_l(\beta^{(0)})$. If $k^{\max} \leq c_0$, defines k_l by $k^{\max}$, choose the group size k_l which minimizes the function $C_l(\beta^{(0)}, k)$, $k_l = \mathrm{argmin}_{k\in\mathscr{K}}C_l(\beta^{(0)}, k)$. Define the set of covariants $G_l = \{\mathbf{x}_{n_{l-1}+1}^{\mathrm{s}}, \ldots, \mathbf{x}_{n_l}^{\mathrm{s}}\}$.

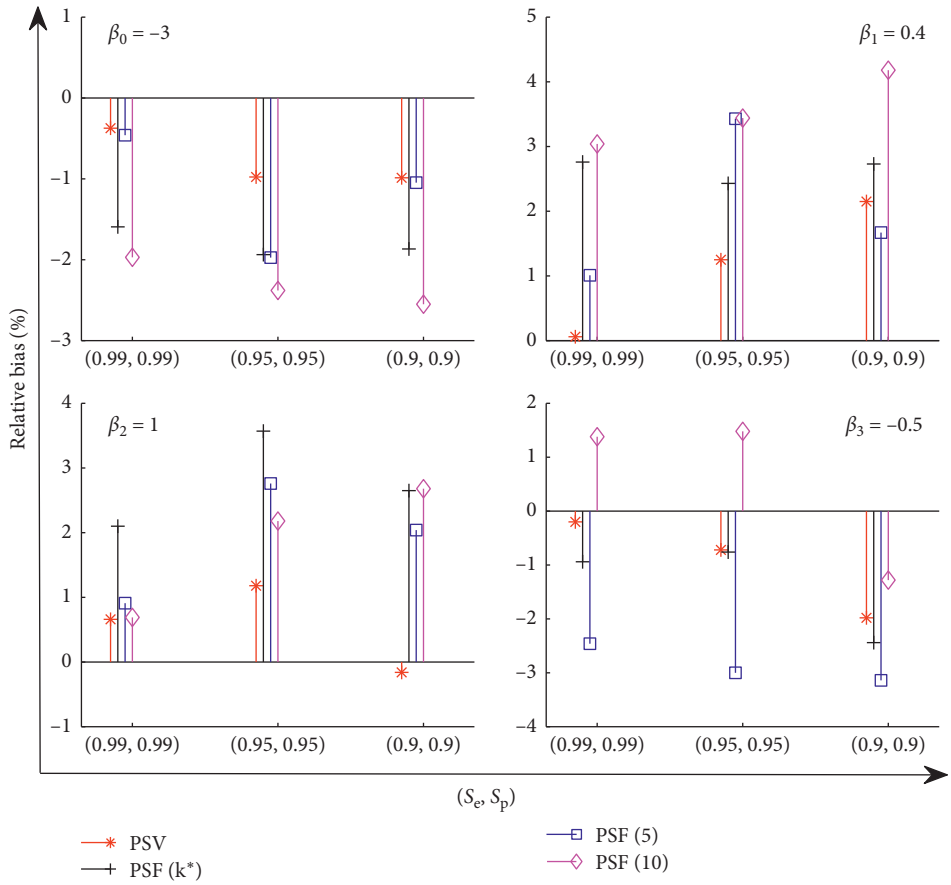

FIGURE 2: The relative bias of the parameters $\beta_0 - \beta_3$ under Model I: $x_1 \sim \Gamma(2.5, 0.8)$, $x_2 \sim B(0.3)$, and $x_3 \sim N(1, 0.5)$, with the number of groups $m = 1000$.

Step 3. Let $G^s = G^s/G_l$, $l = l + 1$. Repeat Step 2 to form the next group in the same way until all specimens are assigned.

Step 4. Screen the groups and obtain maximum likelihood estimator of β.

Note that this is a data-driven pooling strategy. Additionally, the above procedure does not strictly require that all specimens are enrolled before screening since the set G^s is dynamic and could be renewed by new enrolled specimens.

2.3. Numerical Results. In this section, we proceed to evaluate the performance of our proposed procedure. Name it by PSV, which is pooling strategy with varied group sizes. For comparison, we also present the results of pooling strategy with a single group size k, named by PSS(k). The group size k for PSS(k) is given in advance, e.g., $k = 5, 10$, or could be determined by the average prevalence of those enrolled samples. For the latter, we determine the optimal single group size k^* by minimizing the variance of $\widehat{p}$.

To investigate the performance of these methods, define the link function $g(\cdot)$ as the logistic function $g(u) = 1/(1 + \exp(-u))$. Then, individual prevalence is obtained through the following model:

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_0 + \beta_1 x_{1,ij} + \cdots + \beta_{d-1} x_{d-1,ij}, \quad i = 1, \ldots, m, \; j = 1, \ldots, k_i. \tag{4}$$

We first consider a single covariant ($d = 2$), following the normal distribution $N(2, 1.5)$ or the gamma distribution $\Gamma(2.5, 0.8)$. The corresponding parameters are set by $\beta_0 = -3$ and $\beta_1 = 0.4$. The samples are generated under these settings, and the procedures are repeated by $M = 5000$ times. We report the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$, along with their mean square error (MSE) in Table 1 under different settings of sensitivity, specificity, and the number of groups. In Figure 1, we further report the relative bias of the parameters.

Table 1 shows that all procedures have similar performance except PSF [5]. While using the procedure PSF, we have to choose a group size in advance. It is crucial for a group testing algorithm since the precision of estimators severely depend on the group size. In our setting, the average of individual prevalence is about 0.0997, and the corresponding optimal single group size is mostly $k^* = 13, 12, 11$ for $(S_e, S_p) = (0.99, 0.99), (0.95, 0.95)$, and $(0.9, 0.9)$ respectively. Consequently, the procedure PSF [10] has better performance than PSF [5] since the latter procedure uses a

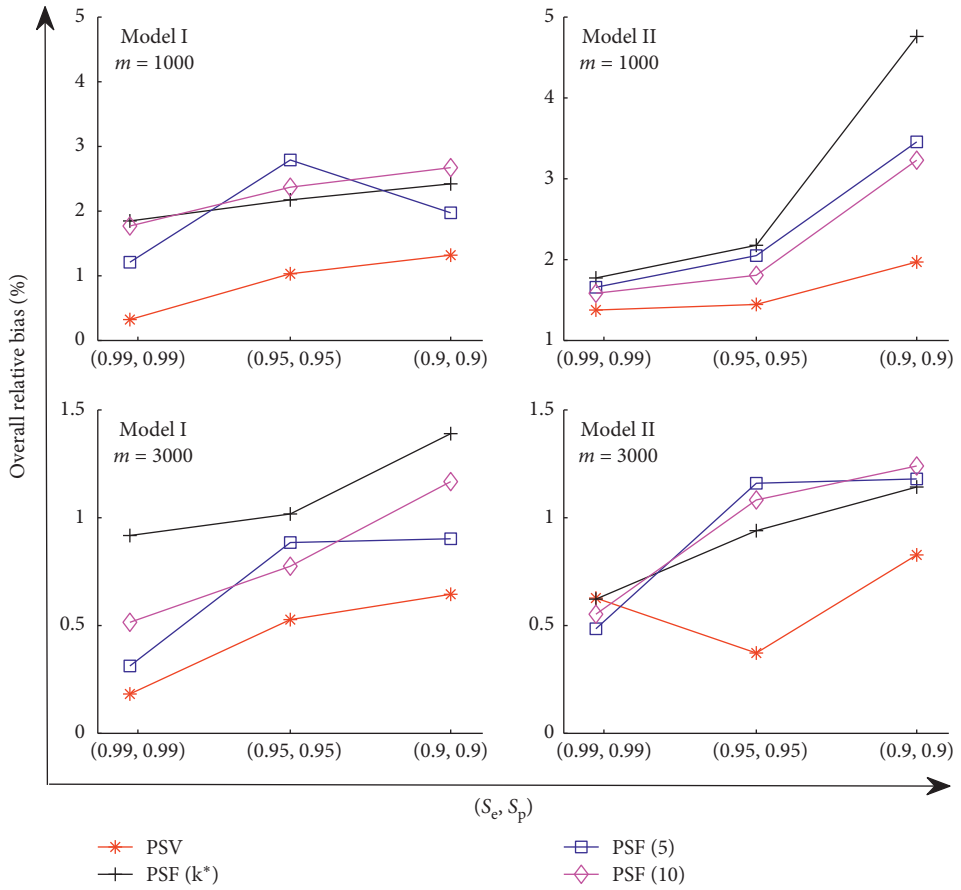

FIGURE 3: The overall relative bias of the parameters, defined as $R = (1/4)\sum_{l=1}^{4}|(\widehat{\beta}_l - \beta_l)/\beta_l|$. Model I: $x_1 \sim \Gamma(2.5, 0.8)$, $x_2 \sim B(0.3)$, and $x_3 \sim N(1, 0.5)$. Model II: $x_1 \sim N(2, 1.5)$, $x_2 \sim B(0.3)$, and $x_3 \sim N(1, 0.5)$.

too smaller group size. Figure 1 further shows the relative bias of the parameters, β_0 and β_1. Our procedure with varying group sizes, PSV, has very good performance under different scenarios. The procedure PSF [5] still has the poorest performance on the measurement of relative bias. As data-driven pooling strategies, PSV and PSF (k^*) both show good performance, but PSV has smaller bias, which is a desired characteristic.

We proceed to consider the model (2) with $d = 4$. Denote the single variable in the above setting by x_1. We add two more variables: x_2 follows the binomial distribution $B(0.3)$ and x_3 follows the normal distribution $N(1, 0.5)$. Then, the model (2) is

$$\text{logit}\left(p_{ij}\right) = \beta_0 + \beta_1 x_{1,ij} + \beta_2 x_{2,ij} + \beta_3 x_{3,ij}, \quad i = 1, \ldots, m,\ j = 1, \ldots, k_i. \tag{5}$$

Specifically, denote by "Model I": $x_1 \sim \Gamma(2.5, 0.8)$, $x_2 \sim B(0.3)$, $x_3 \sim N(1, 0.5)$, and "Model II": $x_1 \sim N(2, 1.5)$, $x_2 \sim B(0.3)$, $x_3 \sim N(1, 0.5)$. Set the parameters by $\beta_0 = -3$, $\beta_1 = 0.4$, $\beta_2 = 1$, and $\beta_3 = -0.5$. In Figure 2, we report the relative bias of the estimators $\widehat{\beta}_0 - \widehat{\beta}_3$ under Model I. Furthermore, define a measurement of $R = (1/4)\sum_{l=1}^{4}|(\widehat{\beta}_l - \beta_l)/\beta_l|$ to calculate the overall relative bias. The results are reported in Figure 3.

Figure 2 shows that our procedure PSV performs best among the four procedures. It is a similar result as shown in Figure 1. The overall relative bias of these estimators reported in Figure 3 also confirms such property. It also reveals that pooling procedures using a single group size are not desired for a heterogeneous population, even the group size is carefully chosen, e.g., k^*.

2.4. An Illustrative Application. Verstraeten et al. conducted a surveillance study in Kenya to monitor a trend in HIV risk over time [43]. The samples were collected from pregnant women, along with potential risk covariants such as age, parity, and education level. They used a common group size of 10 to estimate the seroprevalence of HIV. However, the individual prevalence of HIV is related with those risk covariants, e.g., the risk of HIV might tend to increase with age. For this data set, Vansteelandt et al. reported a set of group sizes varying between 5 and 12 under cost-precision trade-off [40].

We proceed to illustrate our pooling strategy based on part of these data published in [44]. They reported $N = 428$ individuals enrolled in the experiment, including their age (x_1) and education level (x_2). Using model presented in [2], the individual prevalence p_{ij} follows the model: $\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{1,ij} + \beta_2 x_{2,ij}, i = 1, \ldots, m, j = 1, \ldots, k_i$ with $N = \sum_{i=1}^{m} k_i$. Let the initial estimator be $\beta^{(0)} = [-2, -0.05, 0.5]$. Using our proposed pooling strategies PSV and PSF(k^*), the group sizes are listed in Table 2. Correspondingly, we obtain

Table 2: The group sizes chosen using PSV procedure for the Kenyan example.

Procedure	PSV							PSS
Group size	6	7	8	9	10	11	12	11
Number of groups	2	2	2	4	3	4	23	39

estimators: $\widehat{\beta} = [-2.909, -0.033, 0.473]$ using PSV and $\widehat{\beta} = [-3.011, -0.028, 0.443]$ using PSF(k^*).

3. Discussion

In biological and epidemiological studies, there is growing interest in developing methods for a more accurate result but less cost. Group testing is such a cost saving strategy. In this study, we developed a pooling strategy that uses varying group sizes while individual information is available. This strategy is attractive since it only depends on the information of enrolled specimens and does not require a group size chosen in advance. Due to the characteristic of data-driven and theoretical justification, the procedure, "PSV," proposed in this study has a robust performance under different settings. It is convenient for practical application since we do not have to worry about how to choose an appropriate group size.

Varying group sizes are reasonable to be used when the target population is diverse. For example, a sequential testing procedure using several group sizes is adopted to estimate virus infection levels of carnation populations grown in glasshouses since different carnation populations were expected to have a wide range of infection levels [45]. We could pool more specimens into one group if the probability of testing positive is small. It sounds reasonable to balance the probability of testing positive for each group, a way to mimic the situation when all enrolled specimens are homogeneous.

In this study, we also propose a procedure using a single group size k^* determined by minimizing the variance of estimator of the prevalence. We could choose this procedure if we prefer a simple procedure or the diversity among the specimens to be screened is ignorable. Besides, we did not consider the cost of collecting specimens. If a test is much more expensive than that of collecting specimens, then the cost of tests is the main consideration in a project involving large-scale screening. Otherwise, it is necessary to take into account the overall cost of collecting and test while using the pooling strategy.

Data Availability

The Kenya data supporting this study are from previously reported studies and datasets, which have been cited. The data are available at https://cran.r-project.org/package=binGroup.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 11801102 and 11501134), Guangxi Scholarship Fund of Guangxi Education Department, Guangxi Natural Science Foundation (no. 2018GXNSFAA138161), and Research Projects of Guangxi Colleges (no. 2018KY0081).

Supplementary Materials

This article contains additional information on some technical aspects of the research, including the detailed calculation of the Fisher information matrix of the regression parameter and theoretical justification of Step 2 of our sequential adaptive pooling algorithm. (*Supplementary Materials*)

References

[1] F. Behets, S. Bertozzi, M. Kasali et al., "Successful use of pooled sera to determine HIV-1 seroprevalence in Zaire with development of cost-efficiency models," *AIDS*, vol. 4, no. 8, pp. 737–742, 1990.

[2] D. J. Westreich, M. G. Hudgens, S. A. Fiscus, and C. D. Pilcher, "Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests," *Journal of Clinical Microbiology*, vol. 46, no. 5, pp. 1785–1792, 2008.

[3] Z. Zhou, R. M. Mitchell, J. Gutman et al., "Pooled PCR testing strategy and prevalence estimation of submicroscopic infections using bayesian latent class models in pregnant women receiving intermittent preventive treatment at Machinga District Hospital, Malawi, 2010," *Malaria Journal*, vol. 13, no. 1, p. 509, 2014.

[4] D. Leong, K. NicAogáin, L. Luque-Sastre et al., "A 3-year multi-food study of the presence and persistence of Listeria monocytogenes in 54 small food businesses in Ireland," *International Journal of Food Microbiology*, vol. 249, pp. 18–26, 2017.

[5] A. B. Hutchinson, P. Patel, S. L. Sansom et al., "Cost-effectiveness of pooled nucleic acid amplification testing for acute HIV infection after third-generation HIV antibody screening and rapid testing in the United States: a comparison of three public health settings," *PLoS Medicine*, vol. 7, no. 9, article e1000342, 2010.

[6] J. C. Emmanuel, M. T. Bassett, H. J. Smith, and J. A. Jacobs, "Pooling of sera for human immunodeficiency virus (HIV) testing: an economical method for use in developing countries," *Journal of Clinical Pathology*, vol. 41, no. 5, pp. 582–585, 1988.

[7] S. Linauts, J. Saldanha, and D. M. Strong, "PRISM hepatitis B surface antigen detection of hepatits B virus minipool nucleic acid testing yield samples," *Transfusion*, vol. 48, no. 7, pp. 1376–1382, 2008.

[8] P. Mester, A. K. Witte, C. Robben et al., "Optimization and evaluation of the qPCR-based pooling strategy DEP-pooling in dairy production for the detection of Listeria monocytogenes," *Food Control*, vol. 82, pp. 298–304, 2017.

[9] C. Lindan, M. Mathur, S. Kumta et al., "Utility of pooled urine specimens for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR,"

Journal of Clinical Microbiology, vol. 43, no. 4, pp. 1674–1677, 2005.

[10] P. Saha-Chaudhuri and C. R. Weinberg, "Specimen pooling for efficient use of biospecimens in studies of time to a common event," *American Journal of Epidemiology*, vol. 178, no. 1, pp. 126–135, 2013.

[11] E. M. Mitchell, R. H. Lyles, A. K. Manatunga, and E. F. Schisterman, "Semiparametric regression models for a right-skewed outcome subject to pooling," *American Journal of Epidemiology*, vol. 181, no. 7, pp. 541–548, 2015.

[12] R. Dorfman, "The detection of defective members of large populations," *Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[13] J. Tebbs and C. Bilder, "Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 9, no. 1, pp. 79–90, 2004.

[14] J. L. Gastwirth, "The efficiency of pooling in the detection of rare mutations," *American Journal of Human Genetics*, vol. 67, no. 4, pp. 1036–1039, 2000.

[15] M. Ozerov, A. Vasemägi, V. Wennevik et al., "Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative markers for genetic stock identification," *PLoS One*, vol. 8, no. 12, Article ID e82434, 2013.

[16] C. D. Pilcher, M. A. Price, I. F. Hoffman et al., "Frequent detection of acute primary HIV infection in men in Malawi," *AIDS*, vol. 18, no. 3, pp. 517–524, 2004.

[17] S. B. Kim, H. W. Kim, H.-S. Kim et al., "Pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection in Seoul, South Korea," *Scandinavian Journal of Infectious Diseases*, vol. 46, no. 2, pp. 136–140, 2014.

[18] D. H. Seo, D. H. Whang, E. Y. Song et al., "Occult hepatitis B virus infection and blood transfusion," *World Journal of Hepatology*, vol. 7, no. 3, pp. 600–606, 2015.

[19] A. L. Heffernan, L. L. Aylward, L.-M. L. Toms, P. D. Sly, M. Macleod, and J. F. Mueller, "Pooled biological specimens for human biomonitoring of environmental chemicals: opportunities and limitations," *Journal of Exposure Science and Environmental Epidemiology*, vol. 24, no. 3, pp. 225–232, 2014.

[20] M. Ramos, A. L. Heffernan, L. Toms et al., "Concentrations of phthalates and DINCH metabolites in pooled urine from Queensland, Australia," *Environment International*, vol. 88, pp. 179–186, 2016.

[21] W. H. Swallow, "Relative mean squared error and cost considerations in choosing group size for group testing to estimate infection rates and probabilities of disease transmission," *Phytopathology*, vol. 77, no. 10, pp. 1376–1381, 1987.

[22] J. M. Hughes-Oliver and W. H. Swallow, "A two-stage adaptive group-testing procedure for estimating small proportions," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 982–993, 1994.

[23] X. M. Tu, E. Litvak, and M. Pagano, "On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening," *Biometrika*, vol. 82, no. 2, pp. 287–297, 1995.

[24] A. Liu, C. Liu, Z. Zhang, and P. S. Albert, "Optimality of group testing in the presence of misclassification," *Biometrika*, vol. 99, no. 1, pp. 245–251, 2011.

[25] W. Xiong and J. Ding, "Robust procedures for experimental design in group testing considering misclassification," *Statistics & Probability Letters*, vol. 100, pp. 35–41, 2015.

[26] P. Chen, J. M. Tebbs, and C. R. Bilder, "Group testing regression models with fixed and random effects," *Biometrics*, vol. 65, no. 4, pp. 1270–1278, 2009.

[27] Z. Zhang, A. Liu, R. H. Lyles, and B. Mukherjee, "Logistic regression analysis of biomarker data subject to pooling and dichotomization," *Statistics in Medicine*, vol. 31, no. 22, pp. 2473–2484, 2012.

[28] Q. Li, A. Liu, and W. Xiong, "D-optimality of group testing for joint estimation of correlated rare diseases with misclassification," *Statistica Sinica*, vol. 27, no. 2, pp. 823–838, 2017.

[29] J. L. Gastwirth and P. A. Hammick, "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," *Journal of Statistical Planning and Inference*, vol. 22, no. 1, pp. 15–27, 1989.

[30] M. Xie, "Regression analysis of group testing samples," *Statistics in Medicine*, vol. 20, no. 13, pp. 1957–1969, 2001.

[31] C. R. Bilder and J. M. Tebbs, "Bias, efficiency, and agreement for group-testing regression models," *Journal of Statistical Computation and Simulation*, vol. 79, no. 1, pp. 67–80, 2009.

[32] M. Li and M. Xie, "Nonparametric and semiparametric regression analysis of group testing samples," *International Journal of Statistics in Medical Research*, vol. 1, no. 1, pp. 60–72, 2012.

[33] D. Wang, C. S. McMahan, C. M. Gallagher, and K. B. Kulasekera, "Semiparametric group testing regression models," *Biometrika*, vol. 101, no. 3, pp. 587–598, 2013.

[34] A. Delaigle and A. Meister, "Nonparametric regression analysis for group testing data," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 640–650, 2011.

[35] A. Delaigle and W.-X. Zhou, "Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1785–1796, 2015.

[36] C. J. Williams and C. M. Moffitt, "Estimation of fish and wildlife disease prevalence from imperfect diagnostic tests on pooled samples with varying pool sizes," *Ecological Informatics*, vol. 5, no. 4, pp. 273–280, 2010.

[37] G. Hepworth, "Confidence intervals for proportions estimated by group testing with groups of unequal size," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 10, no. 4, pp. 478–497, 2005.

[38] G. Haber and Y. Malinovsky, "Random walk designs for selecting pool sizes in group testing estimation with small samples," *Biometrical Journal*, vol. 59, no. 6, pp. 1382–1398, 2017.

[39] G. Haber, Y. Malinovsky, and P. S. Albert, "Sequential estimation in the group testing problem," *Sequential Analysis*, vol. 37, no. 1, pp. 1–17, 2018.

[40] S. Vansteelandt, E. Goetghebeur, and T. Verstraeten, "Regression models for disease prevalence with diagnostic tests on pools of serum samples," *Biometrics*, vol. 56, no. 4, pp. 1126–1133, 2000.

[41] A. Delaigle and P. Hall, "Nonparametric regression with homogeneous group testing data," *The Annals of Statistics*, vol. 40, no. 1, pp. 131–158, 2012.

[42] C. S. McMahan, J. M. Tebbs, and C. R. Bilder, "Informative dorfman screening," *Biometrics*, vol. 68, no. 1, pp. 287–296, 2012.

[43] T. Verstraeten, B. Farah, L. Duchateau, and R. Matu, "Pooling sera to reduce the cost of HIV surveillance: a feasibility study in a rural Kenyan district," *Tropical Medicine & International Health*, vol. 3, no. 9, pp. 747–750, 1998.

[44] C. R. Bilder, B. Zhang, F. Schaarschmidt, and J. M. Tebbs, "binGroup: a package for group testing," *The R Journal*, vol. 2, no. 2, pp. 56–60, 2010.

[45] G. Hepworth and R. Watson, "Debiased estimation of proportions in group testing," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 58, no. 1, pp. 105–121, 2009.