Web-Based Supplementary Materials for "Prevalence Estimation Subject to Misclassification: The Mis-Substitution Bias and Some Remedies"

Zhiwei Zhang[1,*], Chunling Liu[2], Sungduk Kim[3] and Aiyi Liu[3]

[1]Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland, USA
[2]Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, PR China
[3]Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, USA
*zhiwei.zhang@fda.hhs.gov

Web Appendix A: Information Formulas

We first consider the VS design, where V is allowed to depend on T but not on k, so the sampling mechanism is fully described by $\boldsymbol{\gamma} = (\gamma_1, \gamma_0)$. Let $(\mathrm{Se}, \mathrm{Sp})$ be specified through parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, as in the VGS design. In a group of size k, the Fisher information for $\boldsymbol{\theta} = (\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is given by

$$\mathbf{I}_k = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} & \mathbf{I}_{13} \\ \mathbf{I}_{21} & \mathbf{I}_{22} & \mathbf{I}_{23} \\ \mathbf{I}_{31} & \mathbf{I}_{32} & \mathbf{I}_{33} \end{pmatrix},$$

where

$$\mathbf{I}_{11} = \left(\frac{1-\gamma_1}{p_k} + \frac{1-\gamma_0}{1-p_k}\right)\left(\frac{\partial p_k}{\partial \lambda}\right)^2 + \left(\frac{\mathrm{Se}_k\gamma_1 + (1-\mathrm{Se}_k)\gamma_0}{\pi_k} + \frac{(1-\mathrm{Se}_k)\gamma_1 + \mathrm{Se}_k\gamma_0}{1-\pi_k}\right)\left(\frac{\partial \pi_k}{\partial \lambda}\right)^2,$$

$$\mathbf{I}_{22} = \left(\frac{1-\gamma_1}{p_k} + \frac{1-\gamma_0}{1-p_k}\right)\left(\frac{\partial p_k}{\partial \boldsymbol{\alpha}}\right)^{\otimes 2} + \pi_k\left(\frac{\gamma_1}{\mathrm{Se}_k} + \frac{\gamma_0}{1-\mathrm{Se}_k}\right)\left(\frac{\partial \mathrm{Se}_k}{\partial \boldsymbol{\alpha}}\right)^{\otimes 2},$$

$$\mathbf{I}_{33} = \left(\frac{1-\gamma_1}{p_k} + \frac{1-\gamma_0}{1-p_k}\right)\left(\frac{\partial p_k}{\partial \boldsymbol{\beta}}\right)^{\otimes 2} + (1-\pi_k)\left(\frac{\gamma_1}{1-\mathrm{Sp}_k} + \frac{\gamma_0}{\mathrm{Sp}_k}\right)\left(\frac{\partial \mathrm{Sp}_k}{\partial \boldsymbol{\beta}}\right)^{\otimes 2},$$

$$\mathbf{I}_{21} = \left(\frac{1-\gamma_1}{p_k} + \frac{1-\gamma_0}{1-p_k}\right)\left(\frac{\partial p_k}{\partial \lambda}\right)\left(\frac{\partial p_k}{\partial \boldsymbol{\alpha}}\right) + (\gamma_1 - \gamma_0)\left(\frac{\partial \pi_k}{\partial \lambda}\right)\left(\frac{\partial \mathrm{Se}_k}{\partial \boldsymbol{\alpha}}\right) = \mathbf{I}_{12}^{\mathrm{T}},$$

$$\mathbf{I}_{31} = \left(\frac{1-\gamma_1}{p_k} + \frac{1-\gamma_0}{1-p_k}\right)\left(\frac{\partial p_k}{\partial \lambda}\right)\left(\frac{\partial p_k}{\partial \boldsymbol{\beta}}\right) + (\gamma_1 - \gamma_0)\left(\frac{\partial \pi_k}{\partial \lambda}\right)\left(\frac{\partial \mathrm{Sp}_k}{\partial \boldsymbol{\beta}}\right) = \mathbf{I}_{13}^{\mathrm{T}},$$

$$\mathbf{I}_{23} = \left(\frac{1-\gamma_1}{p_k} + \frac{1-\gamma_0}{1-p_k}\right)\left(\frac{\partial p_k}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial p_k}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right) = \mathbf{I}_{32}^{\mathrm{T}},$$

$\mathrm{Se}_k = \mathrm{Se}(k; \boldsymbol{\alpha})$, $\mathrm{Sp}_k = \mathrm{Sp}(k; \boldsymbol{\beta})$, $\pi_k = \mathrm{P}(D=1) = 1-(1-\pi)^k$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$ for a vector $\boldsymbol{a}$. The information for $(\pi, \mathrm{Se}, \mathrm{Sp})$ is given by $\mathbf{I}_{\mathrm{vs}} = \mathbf{J}^{\mathrm{T}}\mathbf{I}_k\mathbf{J}$, where $\mathbf{J} = \partial\boldsymbol{\theta}/\partial(\pi, \mathrm{Se}, \mathrm{Sp})^{\mathrm{T}}$.

For the VGS design, the Fisher information for $\boldsymbol{\theta}$ is given by $\mathbf{I}_{\mathrm{vgs}} = \sum_{k\in\mathcal{K}} \tau_k \mathbf{I}_{\mathrm{vgs},k}$, where $\mathbf{I}_{\mathrm{vgs},k}$ can be obtained from $\mathbf{I}_k$ (defined above) by setting $\gamma_0 = \gamma_1 = 0$.

Web Appendix B: Nonparametric Bootstrap CIs

Here we describe a general procedure for obtaining bootstrap CIs for an arbitrary parameter θ in the VS or VGS design. In general, the observed data can be represented as $(k_i, T_i, V_i, V_i D_i)$, $i = 1, \ldots, n$. In the VS design, the k_i are identical and the V_i are random. In the VGS design, the k_i are variable and the V_i are identically 0 (so the D_i are never observed). To generate a bootstrap sample, we take a random sample with replacement from $\{1, \ldots, n\}$, and obtain $\{J_1, \ldots, J_n\}$. Then we calculate an estimate of θ based on the boostrap sample $\{(k_{J_i}, T_{J_i}, V_{J_i}, V_{J_i} D_{J_i}), i = 1, \ldots, n\}$. This procedure will be repeated many times, resulting in $\{\widehat{\theta}_b, b = 1, \ldots, B\}$, where B is a large number to be specified by the analyst. Let ξ_p denote the p-quantile of $\{\widehat{\theta}_b, b = 1, \ldots, B\}$. Then a $100(1-\alpha)\%$ confidence interval for θ is obtained as $(\xi_{\alpha/2}, \xi_{1-\alpha/2})$.

Web Appendix C: Identification of $(\pi, \mathrm{Se}, \mathrm{Sp})$ Under the VGS Design and the Constancy Assumption

We assume that the support of k contains three distinct values, say $k_1 < k_2 < k_3$. Let p_k denote the (directly identifiable) probability of a positive test result for a group of size k, given by equation (1). For notational convenience, we write $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, \vartheta_3)$ with $\vartheta_1 = \mathrm{Se}$, $\vartheta_2 = \mathrm{Se} + \mathrm{Sp} - 1$, and $\vartheta_3 = 1 - \pi$. Then equation (1) can be rewritten as

$$p_{k_\ell} = \vartheta_1 - \vartheta_2 \vartheta_3^{k_\ell}, \qquad \ell = 1, 2, 3. \tag{B.1}$$

Simple algebra yields that

$$p_{k_\ell} - p_{k_1} = \vartheta_2 (\vartheta_3^{k_1} - \vartheta_3^{k_\ell}), \qquad \ell = 2, 3, \tag{B.2}$$

and that

$$\frac{p_{k_3} - p_{k_1}}{p_{k_2} - p_{k_1}} = \frac{\vartheta_3^{k_1} - \vartheta_3^{k_3}}{\vartheta_3^{k_1} - \vartheta_3^{k_2}} = \frac{1 - \vartheta_3^{d_{13}}}{1 - \vartheta_3^{d_{12}}}, \tag{B.3}$$

where $d_{1\ell} = k_\ell - k_1$, $\ell = 2, 3$. Now suppose there is another set of parameter values, $\tilde{\boldsymbol{\vartheta}} = (\tilde{\vartheta}_1, \tilde{\vartheta}_2, \tilde{\vartheta}_3)$, satisfying (B.1) and hence (B.2) and (B.3). We will show that $\tilde{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}$, starting with $\tilde{\vartheta}_3 = \vartheta_3$. From equation (B.3) we deduce that

$$\log(1 - \tilde{\vartheta}_3^{d_{13}}) - \log(1 - \tilde{\vartheta}_3^{d_{12}}) = \log(1 - \vartheta_3^{d_{13}}) - \log(1 - \vartheta_3^{d_{12}}).$$

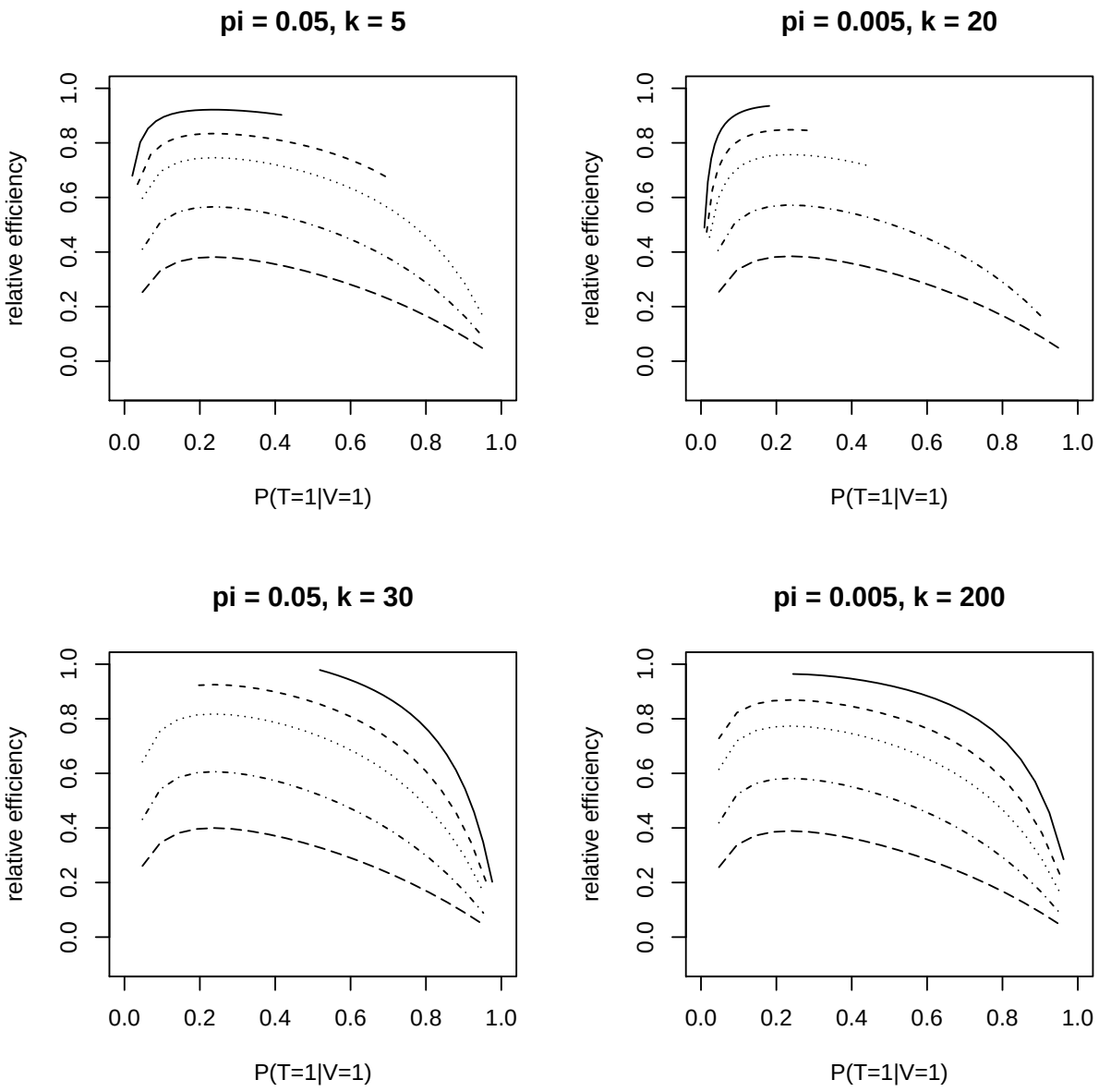
If $\tilde{\vartheta}_3 \neq \vartheta_3$, then the mean value theorem implies that

$$0 = \frac{\mathrm{d}}{\mathrm{d}c} \log \frac{1 - c^{d_{13}}}{1 - c^{d_{12}}} \bigg|_{c = c_0} = \frac{d_{12} c_0^{d_{12}-1}}{1 - c_0^{d_{12}}} - \frac{d_{13} c_0^{d_{13}-1}}{1 - c_0^{d_{13}}}$$

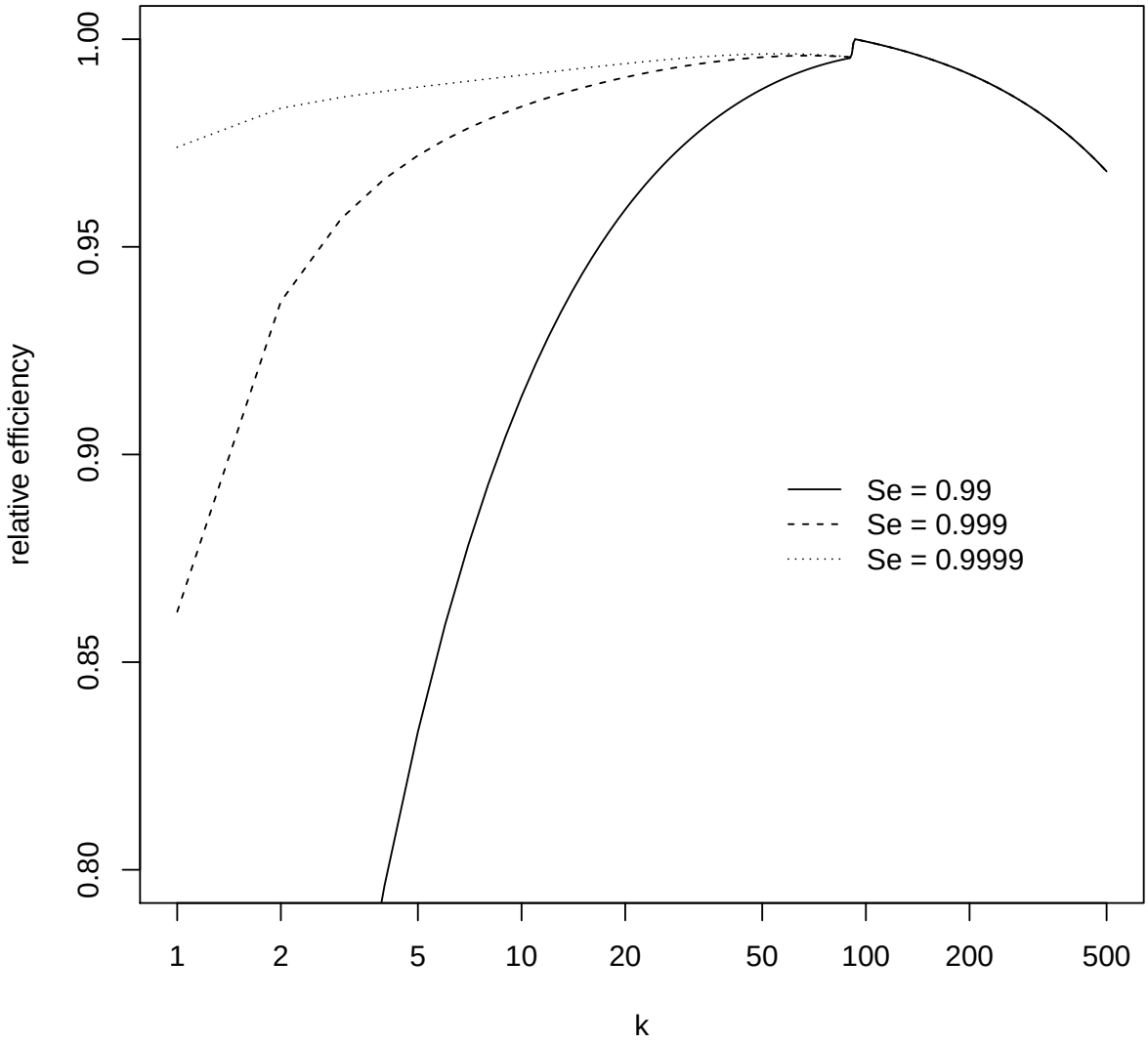
for some c_0 between ϑ_3 and $\tilde{\vartheta}_3$. But this is impossible for $c_0 \in (0, 1)$ and $1 \leq d_{12} < d_{13}$, because the function $f(d) = d c_0^{d-1}/(1 - c_0^d)$ is strictly decreasing. The latter assertion can be verified by taking the derivative:

$$f'(d) = \frac{c_0^{d-1}\{1 + \log(c_0^d) - c_0^d\}}{(1 - c_0^d)^2} < 0.$$

This shows that $\tilde{\vartheta}_3 = \vartheta_3$, which implies $\tilde{\vartheta}_2 = \vartheta_2$ and $\tilde{\vartheta}_1 = \vartheta_1$ through equations (B.2) and (B.1).



Web Figure 1: Allocation of the VS between T-positive and T-negative groups in the absence of a dilution effect: relative efficiency as a function of $p_{k(V)} = \mathrm{P}(T = 1|V = 1)$, the proportion of T-positive groups in the VS, for selected values of π, k, and the sampling fraction $\upsilon = \mathrm{P}(V = 1) \in \{0.5, 0.3, 0.2, 0.1, 0.05\}$ (from top curve to bottom curve within each panel).



Web Figure 2: Design considerations for the Canadian blood donor study: relative efficiency as a function of k, for fixed m and N without a dilution effect.

Web Table 1: Simulation results for the VS design without a dilution effect: empirical relative bias and standard deviation (SD), proportion of replicate samples in which closed-form standard errors are available (SEA) from the observed information matrix, and coverage probability for $(\pi, \text{Se}, \text{Sp})$ among the SEA samples, under different combinations of π, k and the sampling fraction $\upsilon = \text{P}(V = 1)$.

k	υ	Relative Bias			SD (10^{-2})			P(SEA)	P(Coverage|SEA)		
		π	Se	Sp	π	Se	Sp		π	Se	Sp
					$\pi = 0.05$						
1	1	0.001	0.000	0.000	0.22	0.99	0.07	1.00	0.95	0.95	0.95
	0.5	0.000	0.000	0.000	0.22	1.39	0.07	1.00	0.95	0.96	0.95
	0.2	0.000	0.001	0.000	0.25	2.34	0.08	0.98	0.95	0.97	0.95
	0.1	0.000	0.001	0.000	0.28	3.26	0.11	0.87	0.96	0.96	0.95
	0.05	0.000	0.002	0.000	0.35	4.63	0.15	0.63	0.98	0.94	0.96
5	1	0.000	0.000	0.000	0.10	0.45	0.08	1.00	0.95	0.95	0.95
	0.5	0.000	0.000	0.000	0.11	0.64	0.11	1.00	0.95	0.96	0.95
	0.2	0.000	0.000	0.000	0.12	1.01	0.17	1.00	0.95	0.95	0.96
	0.1	0.000	0.000	0.000	0.14	1.40	0.24	0.98	0.95	0.96	0.96
	0.05	0.000	0.000	0.000	0.17	1.98	0.34	0.87	0.95	0.97	0.96
30	1	0.000	0.000	0.000	0.06	0.24	0.15	1.00	0.95	0.95	0.96
	0.5	0.000	0.000	0.000	0.06	0.25	0.26	0.98	0.95	0.95	0.96
	0.2	0.000	0.000	0.000	0.07	0.30	0.60	0.50	0.96	0.95	0.94
	0.1	0.000	0.000	0.000	0.08	0.41	0.85	0.29	0.96	0.95	0.83
	0.05	0.000	0.000	0.000	0.10	0.59	1.16	0.15	0.97	0.94	0.90
60	1	0.000	0.000	0.000	0.07	0.22	0.33	0.90	0.95	0.95	0.96
	0.5	0.000	0.000	0.000	0.07	0.22	0.49	0.65	0.95	0.95	0.96
	0.2	0.000	0.000	0.000	0.07	0.23	0.92	0.25	0.95	0.95	0.87
	0.1	0.001	0.000	0.000	0.08	0.24	1.83	0.06	0.98	0.95	0.00
	0.05	0.001	0.000	0.001	0.10	0.29	2.48	0.03	1.00	0.93	0.00
					$\pi = 0.005$						
1	1	0.001	0.000	0.000	0.07	3.13	0.07	0.91	0.94	0.96	0.95
	0.5	0.002	0.001	0.000	0.07	4.30	0.07	0.71	0.95	0.95	0.95
	0.2	−0.002	0.004	0.000	0.08	6.48	0.07	0.38	0.95	0.87	0.95
	0.1	0.002	0.008	0.000	0.09	8.63	0.07	0.20	0.97	0.88	0.94
	0.05	−0.001	0.017	0.000	0.11	10.41	0.07	0.10	0.97	0.01	0.94
20	1	0.000	0.000	0.000	0.02	0.71	0.07	1.00	0.95	0.95	0.95
	0.5	0.000	0.000	0.000	0.02	1.04	0.08	1.00	0.95	0.95	0.95
	0.2	−0.001	0.001	0.000	0.02	1.65	0.11	1.00	0.95	0.96	0.95
	0.1	0.000	0.001	0.000	0.02	2.35	0.15	0.98	0.95	0.96	0.96
	0.05	0.000	0.001	0.000	0.03	3.31	0.21	0.86	0.97	0.95	0.96
100	1	0.000	0.000	0.000	0.01	0.35	0.09	1.00	0.95	0.95	0.95
	0.5	0.000	0.000	0.000	0.01	0.44	0.16	1.00	0.95	0.95	0.96
	0.2	0.000	0.000	0.000	0.01	0.69	0.25	0.98	0.95	0.95	0.96
	0.1	0.000	0.000	0.000	0.01	0.96	0.36	0.85	0.95	0.95	0.95
	0.05	0.000	0.000	0.000	0.01	1.34	0.51	0.61	0.96	0.95	0.93
200	1	0.000	0.000	0.000	0.01	0.27	0.12	1.00	0.95	0.95	0.95
	0.5	0.000	0.000	0.000	0.01	0.28	0.26	0.98	0.95	0.95	0.97
	0.2	0.000	0.000	0.000	0.01	0.43	0.41	0.76	0.96	0.95	0.93
	0.1	0.000	0.000	0.000	0.01	0.59	0.58	0.51	0.96	0.95	0.93
	0.05	0.000	0.000	0.000	0.01	0.83	0.81	0.30	0.97	0.95	0.84

Web Table 2: Selected k-values in the top candidate VGS designs shown in Figure 6.

π	Constraint	K			
		5	15	50	150
0.05	fixed n	$(1, 3, 5)$	$(1, 6, 15)$	$(1, 15, 50)$	$(1, 22, 150)$
	fixed N	$(1, 2, 5)$	$(1, 4, 15)$	$(1, 7, 50)$	$(1, 12, 150)$
0.005	fixed n	$(1, 3, 5)$	$(1, 6, 15)$	$(1, 15, 50)$	$(1, 22, 150)$
	fixed N	$(1, 2, 5)$	$(1, 4, 15)$	$(1, 7, 50)$	$(1, 12, 150)$

Web Table 3: Simulation results for the VGS design without a dilution effect: empirical bias (relative for π, absolute for the other parameters) and standard deviation (SD), proportion of replicate samples in which closed-form standard errors are available (SEA) from the observed information matrix, and coverage probability for ($\lambda = \text{logit}(\pi), \alpha = \text{logit}(\text{Se}), \beta = \text{logit}(\text{Sp})$) among the SEA samples, under different combinations of π, K and n.

K	n	Bias				SD			P(SEA)			P(Coverage|SEA)		
		π	λ	α	β	λ	α	β	λ	α	β	λ	α	β
						$\pi = 0.05$								
150	1000	−0.01	−0.01	0.15	1.51	0.07	0.97	2.68	0.99	0.99	0.86	0.95	0.97	0.84
50	1000	0.00	−0.01	1.00	1.41	0.10	2.19	2.70	0.98	0.98	0.99	0.89	0.94	0.89
	5000	0.00	0.00	0.40	2.51	0.05	1.44	2.44	1.00	0.94	0.62	0.91	0.97	0.87
	10000	0.00	0.00	0.05	0.83	0.04	0.34	1.96	1.00	1.00	1.00	0.95	0.95	0.91
15	10000	0.02	0.02	1.15	1.03	0.11	2.16	2.03	0.91	0.92	0.98	0.78	0.88	0.93
	100000	0.01	0.01	0.14	1.19	0.05	0.95	2.42	0.90	0.90	0.82	0.92	0.90	0.95
5	100000	0.07	0.06	−0.14	0.48	0.16	1.11	1.21	0.78	0.78	0.89	0.94	0.87	0.94
						$\pi = 0.005$								
150	5000	0.05	0.04	0.98	0.95	0.13	2.38	2.04	0.98	0.98	0.99	0.88	0.84	0.95
	10000	0.03	0.02	1.05	0.58	0.11	2.22	1.45	0.92	0.92	0.96	0.94	0.85	0.95
50	10000	0.29	0.20	−0.19	0.45	0.33	2.11	1.26	0.89	0.89	0.99	0.89	0.73	0.96
	100000	0.05	0.05	−0.01	0.03	0.12	1.17	0.19	0.82	0.82	0.99	0.94	0.86	0.95
15	100000	0.19	0.08	−0.19	0.03	0.34	1.03	0.17	0.57	0.57	0.99	0.92	0.90	0.94
5	100000	0.01	0.00	−0.03	0.03	0.06	0.24	0.19	0.33	0.33	1.00	1.00	1.00	0.96