Statistica Sinica **27** (2017), 823-838
doi:http://dx.doi.org/10.5705/ss.202015.0178

D-OPTIMALITY OF GROUP TESTING FOR JOINT ESTIMATION OF CORRELATED RARE DISEASES WITH MISCLASSIFICATION

Qizhai Li, Aiyi Liu and Wenjun Xiong

Chinese Academy of Sciences, National Institute of Child Health and Human Development and Guangxi Normal University

Abstract: The D-optimal criterion is used to derive an optimality property of group testing in estimation of the prevalence of two rare correlated diseases when the disease statuses are classified with error. Exact ranges of disease prevalence are obtained in which group testing is more efficient than conventional methods of random sampling.

Key words and phrases: Binary outcomes, classification error, D-optimal criterion, group testing, maximum likelihood estimate, prevalence.

1. Introduction

Group (or pooled) testing has proven to be an efficient tool for screening for and estimating the prevalence of a rare disease by testing the pooled specimens and then retesting each subject in the positive group; see, among many others, Dorfman (1943); Graff and Roeloffs (1972); Sobel and Elashoff (1975); Chen and Swallow (1990); Litvak, Tu and Pagano (1994); Tu, Litvak and Pagano (1995); Brookmeyer (1999); Hung and Swallow (1999); Xie et al. (2001); Hepworth and Watson (2009); Bilder, Tebbs and Chen (2010); Delaigle and Meister (2011); McMahan, Tebbs and Bilder (2012); Pritchard and Tebbs (2011); Liu et al. (2012). Due to its effectiveness in reducing time and cost, group testing has been widely used in such areas as HCV or HIV screening (Cahoon-Young et al. (1989); Jarvis et al. (2005)), drug discovery (Gastwirth and Johnson (1994); Kainkaryam and Woolf (2009)), DNA genomic screening (Barcellos et al. (1997)), intellectual disability determination (Chien, Huang and Lung (2009)), arbovirus infection assessment (Gu, Lampman and Novak (2004)), and food contamination detection (Fahey, Ourisson and Degnan (2006)). If one is only interested in estimating the prevalence of a disease, retesting on subjects in the positive groups is not necessary, since the probability of a group being positive is a monotone function of the probability of an individual being positive. This attractive feature makes group testing more appealing in situations where there are limited resources.

There is a sizable literature on disease prevalence estimation using group testing. Most of the literature focuses on a single-disease model, with or without testing errors. In the presence of testing errors, Liu et al. (2012) recently derived conditions under which a group testing strategy is superior to another in yielding more efficient estimation of the disease prevalence. These conditions thus allow practitioners to choose between different group testing strategies.

In many applications, one can encounter a situation where two or more correlated diseases need to be detected simultaneously, using the same assay. For example, chlamydia and gonorrhea, two commonly correlated diseases, were tested on a single assay at the same time to measure their prevalence by using urine specimens collected in the National Health and Nutrition Examination Survey, 1999−2002 (Datta et al. (2007)). Notably, both chlamydia and gonorrhea are detected with error. According to Datta et al. (2007) (and the citations therein), the chlamydia LCx assay has a test sensitivity of approximately 90% to 94% and a specificity of 95% to 98%. The gonorrhea LCx assay incurs a sensitivity of approximately 86% to 92% and specificity greater than 99%.

It is often desirable to simultaneously estimate the prevalence of the diseases. Unlike the single-disease model, however, much less attention has been given to the multiple-disease model with group testing. Here the strategy for proving the optimality of group testing for a single disease cannot be extended directly to multiple diseases, especially when sensitivity and specificity are not equal to 1. Hughes-Oliver and Rosenberger (2000) considered estimating the proportions of individuals with multiple rare traits. Their approach was based on the assumption of no classification error. Recently Tebbs, McMahan and Bilder (2013) proposed a two-stage hierarchical group testing for two correlated diseases with misclassification. Their simulation results suggested that the group testing procedure can be more efficient than individual testing in estimating the diseases' prevalence, an observation consistent with the theory developed in Tu, Litvak and Pagano (1995) and Liu et al. (2012).

To further the results in Liu et al. (2012) and Tebbs, McMahan and Bilder (2013), we aim to investigate the optimality properties of the group testing strategy in estimating the prevalence of two correlated rare diseases whose statuses are classified with errors. We derive conditions under which group testing yields more efficient estimation than individual testing and other conventional testing strategies.

2. D-Optimality in Group Testing

2.1. General framework

Suppose that n subjects are enrolled and their plasma are collected to evaluate the prevalence of two correlated diseases (D_1 and D_2). Let the disease prevalence be $p_{ab} = \Pr(D_1 = a, D_2 = b)$, $a, b \in \{0, 1\}$, where $D_1 = 1$ means that an individual has the disease D_1 and 0 otherwise, and similarly D_2. To estimate p_{00}, p_{10}, p_{01} and p_{11}, group testing is used with the plasma of k individuals being pooled in one group and assayed. Without loss of generality, we assume that $n = mk$. Let $(X, Y)^\tau$ denote the real status of $(D_1, D_2)^\tau$ in a group, where τ stands for the transpose of a vector or a matrix. We take X=1 (or 0) and $Y = 1$ (or 0) to mean that at least one (or no) subject in the group has D_1 and at least one (or no) subject has D_2. Let $(\tilde{X}, \tilde{Y})^\tau$ be the testing result of the group from the assay. We assume that the statuses of D_1 and D_2 are measured with error, and write the error probability for $(a, b) \neq (c, d)$ as

$$\pi_{ab}^{cd} = \Pr\left(\tilde{X} = c, \tilde{Y} = d \middle| X = a, Y = b\right),\ a, b, c, d \in \{0, 1\}.$$

Let $q_{ab} = \Pr(X = a, Y = b)$, $a, b = 0, 1$. Then, $q_{00} = p_{00}^k$, $q_{01} = (p_{00} + p_{01})^k - p_{00}^k$, $q_{10} = (p_{00} + p_{10})^k - p_{00}^k$, $q_{11} = 1 + p_{00}^k - (p_{00} + p_{10})^k - (p_{00} + p_{01})^k$ and $g_{ab} = \Pr(\tilde{X} = a, \tilde{Y} = b) = \pi_{00}^{ab} q_{00} + \pi_{10}^{ab} q_{10} + \pi_{01}^{ab} q_{01} + \pi_{11}^{ab} q_{11},\quad a, b \in \{0, 1\}$.

Let m_{ab} be the number of groups with assay results (a, b). According to Tu, Litvak and Pagano (1995) and Liu et al. (2012), then g_{ab} satisfies $\pi_{a^- b^-}^{ab} \leq g_{ab} \leq \pi_{ab}^{ab}$, and the maximum likelihood estimate of g_{ab} is

$$\hat{g}_{ab} = \min\left\{\pi_{ab}^{ab}, \max\left\{\pi_{a^- b^-}^{ab}, \frac{m_{ab}}{m}\right\}\right\}, a^- = 1 - a,\ b^- = 1 - b.$$

Once we have $\hat{g}_{ab}$, we can obtain the maximum likelihood estimates $\hat{p}_{cd}$ of p_{cd}, $a, b, c, d \in \{0, 1\}$. Technical details are given in Appendix A. We parameterize the notations as

$$\pi = \left(\pi_{00}^{00}, \pi_{10}^{00}, \pi_{01}^{00}, \pi_{11}^{00}, \pi_{00}^{10}, \pi_{10}^{10}, \pi_{01}^{10}, \pi_{11}^{10}, \pi_{00}^{01}, \pi_{10}^{01}, \pi_{01}^{01}, \pi_{11}^{01}, \pi_{00}^{11}, \pi_{10}^{11}, \pi_{01}^{11}, \pi_{11}^{11}\right)^\tau,$$

$\mathbf{p} = (p_{00}, p_{10}, p_{01})^\tau$, $\hat{\mathbf{p}} = (\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01})^\tau$, and $\mathbf{g} = (g_{00}, g_{10}, g_{01})^\tau$. Since $p_{11} = 1 - p_{00} - p_{10} - p_{01}$, such a definition of the parameter $\mathbf{p} = (p_{00}, p_{01}, p_{10})^\tau$ enables us to reduce the dimension of the parameter space, following Tebbs, McMahan and Bilder (2013). Then the covariance matrix of $\hat{\mathbf{p}}$ is $V_{\mathbf{p}} = I_{\mathbf{p}}^{-1}(\mathbf{p})$, whose derivation is given in Appendix B, where $I_{\mathbf{p}}(\mathbf{p}) = \left(\frac{\partial \mathbf{g}}{\partial \mathbf{q}} \frac{\partial \mathbf{q}}{\partial \mathbf{p}}\right)^\tau I_{\mathbf{g}}(\mathbf{g}) \left(\frac{\partial \mathbf{g}}{\partial \mathbf{q}} \frac{\partial \mathbf{q}}{\partial \mathbf{p}}\right)$,

$$I_{\mathbf{g}}(\mathbf{g}) = m \begin{pmatrix} \frac{(g_{00}+g_{11})}{g_{00} g_{11}} & \frac{1}{g_{11}} & \frac{1}{g_{11}} \\ \frac{1}{g_{11}} & \frac{(g_{10}+g_{11})}{g_{10} g_{11}} & \frac{1}{g_{11}} \\ \frac{1}{g_{11}} & \frac{1}{g_{11}} & \frac{(g_{01}+g_{11})}{g_{01} g_{11}} \end{pmatrix},\ \frac{\partial \mathbf{g}}{\partial \mathbf{q}} = \begin{pmatrix} \pi_{00}^{00} - \pi_{11}^{00} & \pi_{10}^{00} - \pi_{11}^{00} & \pi_{01}^{00} - \pi_{11}^{00} \\ \pi_{00}^{10} - \pi_{11}^{10} & \pi_{10}^{10} - \pi_{11}^{10} & \pi_{01}^{10} - \pi_{11}^{10} \\ \pi_{00}^{01} - \pi_{11}^{01} & \pi_{10}^{01} - \pi_{11}^{01} & \pi_{01}^{01} - \pi_{11}^{01} \end{pmatrix},$$

$$\frac{\partial \mathbf{q}}{\partial \mathbf{p}} = k p_{00}^{k-1} \begin{pmatrix} 1 & 0 & 0 \\ (1 + \frac{p_{10}}{p_{00}})^{k-1} - 1 & (1 + \frac{p_{10}}{p_{00}})^{k-1} & 0 \\ (1 + \frac{p_{01}}{p_{00}})^{k-1} - 1 & 0 & (1 + \frac{p_{01}}{p_{00}})^{k-1} \end{pmatrix}.$$

While there are different ways to evaluate the efficiency of an estimate of a parameter vector, we adopt the D-optimal criterion (Kiefer (1974)) that minimizes the determinant of $V_{\mathbf{p}}$, written as $\det(V_{\mathbf{p}})$. With corresponding probabilities given, we seek a group size that minimizes $\det(V_{\mathbf{p}})$.

Theorem 1. *If $\psi(k, \pi, \mathbf{p}) = det\big(cov(V_{\mathbf{p}})\big)$, then for fixed π and $\mathbf{p}$, there exists a k_{opt} that minimizes $\psi(k, \pi, \mathbf{p})$.*

For a proof, see Appendix C. From Theorem 1, there exists an optimal k_{opt} that minimizes $\psi(k, \pi, \mathbf{p})$. k_{opt} can be obtained by minimizing $\det\big(\mathrm{cov}(V_{\mathbf{p}})\big)$ using the Newton-Raphson algorithm since it is a one-dimensional optimization. If $k_{opt} = 1$, then the D-Optimal design reduces to individual testing.

2.2. Optimality of group testing

For a single rare disease, Liu et al. (2012) obtained conditions for group testing to be more efficient than random sampling with a fixed number of groups or a fixed number of subjects. It is desirable to derive such conditions for two correlated rare diseases. Based on Theorem 1, we find that a necessary and sufficient condition that group testing is more efficient than random sampling is $\psi(2, \pi, \mathbf{p}) < \psi(1, \pi, \mathbf{p})$, using the D-optimal criterion. Theorems 2 and 3 give the results that group testing improves the precision in estimating $\mathbf{p}$. Details are given in Appendix D.

As notation, we write $\xi_{00} = \pi_{10}^{00} + \pi_{01}^{00} + 2\pi_{11}^{00}$, $\eta_{00} = \pi_{10}^{00} + \pi_{01}^{00}$, $\zeta_{00} = \pi_{00}^{00}$, $\xi_{10} = \pi_{10}^{10} + \pi_{01}^{10} + 2\pi_{11}^{10}$, $\eta_{10} = \pi_{10}^{10} + \pi_{01}^{10}$, $\zeta_{10} = \pi_{00}^{10}$, $\xi_{01} = \pi_{10}^{01} + \pi_{01}^{01} + 2\pi_{11}^{01}$, $\eta_{01} = \pi_{10}^{01} + \pi_{01}^{01}$, $\zeta_{01} = \pi_{00}^{01}$, $\xi_{11} = \pi_{10}^{11} + \pi_{01}^{11} + 2\pi_{11}^{11}$, $\eta_{11} = \pi_{11}^{11}$, $\zeta_{11} = \pi_{00}^{11}$. We invoke one condition on the joint probabilities, a natural requirement for a screening tool to be practically useful.

Condition: $\pi_{a^-b^-}^{ab} < \min\{\pi_{a^-b}^{ab}, \pi_{ab^-}^{ab}\}$, $\max\{\pi_{ab^-}^{ab}, \pi_{a^-b}^{ab}\} < \pi_{ab}^{ab}$.

Theorem 2 (Fixed number of groups)**.** *Suppose the number of groups is fixed at m, and let δ_1 be the solution of the equation*

$$\frac{p_{00}^2}{2^6(p_{00}+p_{10})^2(p_{00}+p_{01})^2} \prod_{a,b\in\{0,1\}} \frac{\xi_{ab}\delta^2 + 2\eta_{ab}\delta + \zeta_{ab}}{\eta_{ab}\delta + \zeta_{ab}} = 1$$

for δ. If $\delta_1 > \max\big\{p_{10}/p_{00} + p_{11}/p_{00}, p_{01}/p_{00} + p_{11}/p_{00}\big\}$, then group testing with size k is more efficient in estimating the disease prevalence $\mathbf{p}$ than a random sample of size m under the D-optimal criterion.

Table 1. Values of δ_1 and δ_2.

$\mathbf{p}$	δ_1	δ_2	$c_{\mathbf{p}}$
(0.901,0.079,0.019)	1.1586	0.0926	0.0888
(0.916,0.079,0.004)	1.1568	0.0919	0.0873
(0.931,0.049,0.019)	1.1577	0.0923	0.0537
(0.941,0.049,0.009)	1.157	0.092	0.0531
(0.966,0.014,0.019)	1.1567	0.0919	0.0207
(0.971,0.019,0.009)	1.1566	0.0918	0.0206
(0.976,0.009,0.014)	1.1565	0.0918	0.0154
(0.981,0.009,0.009)	1.1565	0.0918	0.0102
(0.896,0.099,0.004)	1.1569	0.092	0.1116
(0.879,0.119,0.001)	1.1565	0.0918	0.1365

Theorem 3 (Fixed number of subjects)**.** *Suppose the total of subjects is fixed at n, and let δ_2 be the solution of the equation*

$$\frac{p_{00}^2}{2^3(p_{00}+p_{10})^2(p_{00}+p_{01})^2}\prod_{a,b\in\{0,1\}}\frac{\xi_{ab}\delta^2+2\eta_{ab}\delta+\zeta_{ab}}{\eta_{ab}\delta+\zeta_{ab}}=1$$

for δ. If $\delta_2>\max\{p_{10}/p_{00}+p_{11}/p_{00},p_{01}/p_{00}+p_{11}/p_{00}\}$, then group testing with size k and m groups is more efficient in estimating the disease prevalence $\mathbf{p}$ than a random sample of size n under the D-optimal criterion.

3. Numerical Results

3.1. Selection of δ_1 and δ_2

To gain some insight into the range of $\mathbf{p}$ where group testing is more efficient than random sampling, we conducted numerical studies to obtain the values of δ_1 and δ_2, for practical values of testing error rates. Let the error rate be $\pi^\tau=(0.96,0.06,0.1,0.006,0.03,0.9,0.004,0.1,0.009,0.008,0.89,0.03,0.001,0.032,0.006,0.864)$. The prevalence of the two correlated diseases were chosen from the set $\{0.01,0.015,0.02,0.05,0.08\}$ and the joint prevalence was 0.001. We also calculated $\max\{p_{10}/p_{00}+p_{11}/p_{00},p_{01}/p_{00}+p_{11}/p_{00}\}$, denoted by $c_{\mathbf{p}}$. Table 1 presents the numerical results. The last two rows of Table 1 correspond to situations in which δ_1 and δ_2 are not larger than $c_{\mathbf{p}}$. They indicate that the conditions in Theorems 2 and 3 are satisfied. For example, when $\mathbf{p}=(0.966,0.014,0.019)$, $\delta_1=1.1567$, and $\delta=0.0919$, both are larger than 0.0207.

Denote efficiency by $\mathrm{EFF}(k)=\psi(1,\pi,\mathbf{p})/\psi(k,\pi,\mathbf{p})$. Here, we consider the case that the subject number is fixed. We present the results for different group sizes and highlight the largest value of efficiency using the settings as above. The

Table 2. Efficiency of group testing compared with individual testing.

	the group size k							
p	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00
(0.901,0.079,0.019)	**1.02**	0.83	0.65	0.51	0.40	0.32	0.25	0.20
(0.916,0.079,0.004)	1.64	**1.81**	1.76	1.62	1.45	1.28	1.11	0.96
(0.931,0.049,0.019)	**1.25**	1.14	0.97	0.81	0.68	0.57	0.48	0.41
(0.941,0.049,0.009)	1.59	**1.70**	1.62	1.48	1.32	1.18	1.04	0.92
(0.966,0.014,0.019)	1.99	2.40	**2.51**	2.45	2.33	2.18	2.03	1.87
(0.971,0.019,0.009)	2.21	2.86	3.14	**3.22**	3.17	3.07	2.93	2.77
(0.976,0.009,0.014)	2.43	3.38	3.91	4.17	**4.25**	4.22	4.13	4.00
(0.981,0.009,0.009)	2.69	4.01	4.88	5.42	5.72	5.86	**5.88**	5.84
(0.896,0.099,0.004)	1.50	**1.56**	1.45	1.28	1.09	0.92	0.77	0.63
(0.879,0.119,0.001)	1.70	1.93	**1.90**	1.73	1.52	1.29	1.07	0.87

corresponding group size is the optimal size to be chosen. We also calculated EFF(k). Table 2 shows the results. It indicates that group testing is more efficient than individual testing even when the Condition is not satisfied. For example, when $\mathbf{p} = (0.971, 0.019, 0.009)$, the optimal group size is 5 with the EFF$(5) = 3.22$.

Both cases reveal the common feature that group testing gains efficiency as compared to random sampling when the conditions of Theorems 2 and 3 hold.

3.2. An example: Estimation of Chlamydia and Gonorrhea prevalence

Chlamydia is the most common sexually transmitted bacterial infection, affecting 3-4 million people each year in the US. It causes pelvic inflammatory disease, ectopic pregnancy, and infertility in women, and testicular and prostate infections, and sterility in men. Gonorrhea is a bacterial infection that often co-exists with chlamydia.

Due to the low prevalence and the serious consequences of the diseases, a group testing strategy is deemed desirable to simultaneously assay-test Chlamydia and Gonorrhea. To demonstrate how to choose the group size in such a study with the methods proposed here, we use data from Datta et al. (2007). In the National Health and Nutrition Examination Survey from 1992 to 2002, 6,632 urine specimens were collected in the National Health and assayed in Abbott Laboratories, Abbott Park, Illinois.

Based on the information from the Introduction, we assumed that the sensitivity and specificity are about 92% and 96.5% for screening chlamydia and about 89% and 99% for screening gonorrhea, using urine specimens with LCx assay. From previous studies we assumed that the individual prevalence of gonorrheal and chlamydial infections was 0.24% and 2.2%, respectively, and joint prevalence

Table 3. Efficiency EFF(k) for different group sizes.

k	1	2	3	4	5	6	7	8
EFF(k)	1.00	3.15	5.27	7.04	8.44	9.49	10.25	10.78
k	9	10	11	12	13	14	15	
EFF(k)	11.11	11.30	11.37	11.34	11.24	11.08	10.88	

of gonorrhea and chlamydia was 0.01104%. These yield π^τ= (0.95535,0.0792, 0.10615, 0.0088, 0.03465, 0.9108, 0.00385, 0.1012, 0.00965, 0.0008, 0.85885, 0.0712, 0.00035, 0.0092, 0.03115, 0.8188).

Suppose a study is planned in order to update the prevalence of the diseases. Checking the conditions in Theorems 2 and 3, we find that $\delta_1 = 1.111$ and $\delta_2 = 0.095$ and $\max\{(p_{10}+p_{11})/p_{00}, (p_{01}+p_{11})/p_{00}\} = 0.0225$, the conditions are satisfied. Fixing the total number of subjects to be 6,632, the D-optimal group size is found to be $k_{opt} = 11$ and the optimal number of groups is $m = 603$. More explicitly, we present the efficiency EFF(k) in Table 3.

This group testing design not only yields more precise prevalence estimation, but is also cost effective. Assuming that the cost per test assay is $16, the cost of $106,112 from individual testing reduces to $9,648 using such a group testing strategy.

4. Discussion

In the present paper, we derived conditions on group sizes such that group testing is more efficient than individual testing in simultaneously estimating the prevalence of two related rare diseases, thereby filling some gaps in the literature. Given empirical knowledge about these prevalences, and the error rates of the test assay, the conditions derived provide useful guideline for designing prospective studies to update the estimation of prevalence. If no prior knowledge is available, one can take an adaptive approach similar to those in Hughes-Oliver and Swallow (1994) and Ridout (1995), by first obtaining estimates of the prevalence and error rates using interim data and then redesigning the study for the remaining group testing with the updated group size.

Taking a look at the conditions in Theorems 2 and 3, when the diseases are rare, $\max\left\{p_{10}/p_{00} + p_{11}/p_{00}, p_{01}/p_{00} + p_{11}/p_{00}\right\} \approx \max\{p_{10}, p_{01}\} + p_{11}$, which is very small. Although there are no closed forms for δ_1 and δ_2, simulation studies show that $\hat{\delta}_1$ and $\hat{\delta}_2$ are most likely large, and are much larger than $\max\{p_{10}, p_{01}\} + p_{11}$, in which case the conditions hold for rare diseases.

Acknowledgement

The authors thank an associate editor and two anonymous reviewers for their insightful comments that improved the manuscript. Q. Li was supported in part by the National Science Foundation of China, Grant No. 11371353, 61134013 and the Breakthrough Project of Strategic Priority Program of the Chinese Academy of Sciences, Grant No. XDB13040600. Research of A. Liu was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH). Research of W. Xiong was supported by Guangxi Natural Science Foundation (2015GXNSFBA139016).

Appendix A. Maximum Likelihood Estimates

The maximum likelihood estimate of p_{ab}, denoted by $\hat{p}_{ab}$, $a, b \in \{0, 1\}$ can be obtained by solving the equations,

$$\begin{cases} \hat{g}_{00} = \pi^{00}_{00} p^k_{00} + \pi^{00}_{10}[(p_{00}+p_{10})^k - p^k_{00}] + \pi^{00}_{01}[(p_{00}+p_{01})^k - p^k_{00}] \\ \qquad +\pi^{00}_{11}[1 + p^k_{00} - (p_{00}+p_{10})^k - (p_{00}+p_{01})^k], \\ \hat{g}_{01} = \pi^{01}_{00} p^k_{00} + \pi^{01}_{10}[(p_{00}+p_{10})^k - p^k_{00}] + \pi^{01}_{01}[(p_{00}+p_{01})^k - p^k_{00}] \\ \qquad +\pi^{01}_{11}[1 + p^k_{00} - (p_{00}+p_{10})^k - (p_{00}+p_{01})^k], \\ \hat{g}_{10} = \pi^{10}_{00} p^k_{00} + \pi^{10}_{10}[(p_{00}+p_{10})^k - p^k_{00}] + \pi^{10}_{01}[(p_{00}+p_{01})^k - p^k_{00}] \\ \qquad +\pi^{10}_{11}[1 + p^k_{00} - (p_{00}+p_{10})^k - (p_{00}+p_{01})^k]. \end{cases}$$

Let

$$A = \begin{vmatrix} \hat{g}_{00} & \pi^{00}_{10} - \pi^{00}_{11} & \pi^{00}_{01} - \pi^{00}_{11} \\ \hat{g}_{01} & \pi^{01}_{10} - \pi^{01}_{11} & \pi^{01}_{01} - \pi^{01}_{11} \\ \hat{g}_{10} & \pi^{10}_{10} - \pi^{10}_{11} & \pi^{10}_{01} - \pi^{10}_{11} \end{vmatrix},$$

$$B = \begin{vmatrix} \pi^{00}_{00} - \pi^{00}_{10} - \pi^{00}_{01} + \pi^{00}_{11} & \hat{g}_{00} & \pi^{00}_{01} - \pi^{00}_{11} \\ \pi^{01}_{00} - \pi^{01}_{10} - \pi^{01}_{01} + \pi^{01}_{11} & \hat{g}_{01} & \pi^{01}_{01} - \pi^{01}_{11} \\ \pi^{10}_{00} - \pi^{10}_{10} - \pi^{10}_{10} + \pi^{10}_{11} & \hat{g}_{10} & \pi^{10}_{01} - \pi^{10}_{11} \end{vmatrix},$$

$$C = \begin{vmatrix} \pi^{00}_{00} - \pi^{00}_{10} - \pi^{00}_{01} + \pi^{00}_{11} & \pi^{00}_{10} - \pi^{00}_{11} & \hat{g}_{00} \\ \pi^{01}_{00} - \pi^{01}_{10} - \pi^{01}_{01} + \pi^{01}_{11} & \pi^{01}_{10} - \pi^{01}_{11} & \hat{g}_{01} \\ \pi^{10}_{00} - \pi^{10}_{10} - \pi^{10}_{10} + \pi^{10}_{11} & \pi^{10}_{10} - \pi^{10}_{11} & \hat{g}_{10} \end{vmatrix},$$

$$D = \begin{vmatrix} \pi^{00}_{00} - \pi^{00}_{10} - \pi^{00}_{01} + \pi^{00}_{11} & \pi^{00}_{10} - \pi^{00}_{11} & \pi^{00}_{01} - \pi^{00}_{11} \\ \pi^{01}_{00} - \pi^{01}_{10} - \pi^{01}_{01} + \pi^{01}_{11} & \pi^{01}_{10} - \pi^{01}_{11} & \pi^{01}_{01} - \pi^{01}_{11} \\ \pi^{10}_{00} - \pi^{10}_{10} - \pi^{10}_{10} + \pi^{10}_{11} & \pi^{10}_{10} - \pi^{10}_{11} & \pi^{10}_{01} - \pi^{10}_{11} \end{vmatrix}.$$

Then $\hat{p}_{00} = (A/D)^{1/k}$, $\hat{p}_{10} = (B/D)^{1/k} - (A/D)^{1/k}$, $\hat{p}_{01} = (C/D)^{1/k} - (A/D)^{1/k}$, and $\hat{p}_{11} = 1 - \hat{p}_{00} - \hat{p}_{10} - \hat{p}_{01}$.

Appendix B. Derivation of $I_{\mathbf{g}}(\mathbf{g})$

Using the notation in the main text, we have

$$\begin{aligned} g_{ab} &= \pi^{ab}_{ab} q_{ab} + \pi^{ab}_{a^-b} q_{a^-b} + \pi^{ab}_{ab^-} q_{ab^-} + \pi^{ab}_{a^-b^-}(1 - q_{ab} - q_{a^-b} - q_{ab^-}) \\ &= \pi^{ab}_{a^-b^-} + (\pi^{ab}_{ab} - \pi^{ab}_{a^-b^-}) q_{ab} + (\pi^{ab}_{a^-b} - \pi^{ab}_{a^-b^-}) q_{a^-b} + (\pi^{ab}_{ab^-} - \pi^{ab}_{a^-b^-}) q_{ab^-}. \end{aligned}$$

The likelihood function is

$$L(\mathbf{g}) = g_{00}^{m_{00}} g_{10}^{m_{10}} g_{01}^{m_{01}} (1 - g_{00} - g_{10} - g_{01})^{m_{11}}$$

and the log-likelihood function is

$$l(\mathbf{g}) = m_{00} \ln g_{00} + m_{10} \ln g_{10} + m_{01} \ln g_{01} + m_{11} \ln(1 - g_{00} - g_{10} - g_{01}).$$

Then,

$$\begin{aligned} -E\Big(\frac{\partial^2 l}{\partial g_{ab}^2}\Big) &= E\left[\frac{m_{ab}}{g_{ab}^2} - \frac{m_{11}}{(1 - g_{00} - g_{10} - g_{01})^2}\right] = \frac{m(g_{ab} + g_{11})}{g_{ab} g_{11}}, \\ -E\left(\frac{\partial^2 l}{\partial g_{ab} \partial g_{cd}}\right) &= E\left[\frac{m_{11}}{(1 - g_{00} - g_{10} - g_{01})^2}\right] = \frac{m}{g_{11}}, (a, b) \neq (c, d). \end{aligned}$$

So,

$$I_{\mathbf{g}}(\mathbf{g}) = m \begin{pmatrix} \frac{(g_{00}+g_{11})}{g_{00}g_{11}} & \frac{1}{g_{11}} & \frac{1}{g_{11}} \\ \frac{1}{g_{11}} & \frac{(g_{10}+g_{11})}{g_{10}g_{11}} & \frac{1}{g_{11}} \\ \frac{1}{g_{11}} & \frac{1}{g_{11}} & \frac{(g_{01}+g_{11})}{g_{01}g_{11}} \end{pmatrix}.$$

Appendix C. Proof of Theorem 1

Let

$$\begin{aligned} \psi_1(k, \pi, \mathbf{p}) &= \frac{g_{00} g_{01} g_{10} g_{11}}{m^3 k^6 p_{00}^{2k} (p_{00} + p_{10})^{2k} (p_{00} + p_{01})^{2k}} \\ &= \frac{p_{00}^{2k}}{m^3 k^6 (p_{00} + p_{10})^{2k} (p_{00} + p_{01})^{2k}} \frac{g_{00}}{p_{00}^k} \frac{g_{10}}{p_{00}^k} \frac{g_{01}}{p_{00}^k} \frac{g_{11}}{p_{00}^k}. \end{aligned}$$

Then $\psi(k, \pi, \mathbf{p}) = \begin{vmatrix} \pi^{00}_{00} - \pi^{00}_{11} & \pi^{00}_{10} - \pi^{00}_{11} & \pi^{00}_{01} - \pi^{00}_{11} \\ \pi^{10}_{00} - \pi^{10}_{11} & \pi^{10}_{10} - \pi^{10}_{11} & \pi^{10}_{01} - \pi^{10}_{11} \\ \pi^{01}_{00} - \pi^{01}_{11} & \pi^{01}_{10} - \pi^{01}_{11} & \pi^{01}_{01} - \pi^{01}_{11} \end{vmatrix}^{-2} \times \psi_1(k, \pi, \mathbf{p})$. To show g_{ab} depends on k, we use $g_{ab}(k)$ as a surrogate of g_{ab}.

For $a, b \in \{0,1\}$, we have

$$\frac{g_{ab}(k)}{p_{00}^k} = \pi_{00}^{ab} + \pi_{10}^{ab}\Big[\Big(1+\frac{p_{10}}{p_{00}}\Big)^k - 1\Big] + \pi_{01}^{ab}\Big[\Big(1+\frac{p_{01}}{p_{00}}\Big)^k - 1\Big] + \pi_{11}^{ab}\Big[1 + p_{00}^{-k} - \Big(1+\frac{p_{10}}{p_{00}}\Big)^k - \Big(1+\frac{p_{01}}{p_{00}}\Big)^k\Big],$$

$$\begin{aligned} &1 + p_{00}^{-k} - \Big(1+\frac{p_{10}}{p_{00}}\Big)^k - \Big(1+\frac{p_{01}}{p_{00}}\Big)^k \\ &= 1 + \exp\Big\{k \ln \frac{1}{p_{00}}\Big\} - \exp\Big\{k \ln\Big(1+\frac{p_{10}}{p_{00}}\Big)\Big\} - \exp\Big\{k \ln\Big(1+\frac{p_{01}}{p_{00}}\Big)\Big\} \\ &= \sum_{i=1}^{+\infty} \frac{\ln^i(\frac{1}{p_{00}}) - \ln^i(1+\frac{p_{10}}{p_{00}}) - \ln^i(1+\frac{p_{01}}{p_{00}})}{i!} k^i, \end{aligned} \tag{C.1}$$

$$\begin{aligned} \frac{p_{00}^{2k}}{(p_{00}+p_{10})^{2k}(p_{00}+p_{01})^{2k}} &= \exp(2(\ln(p_{00}) - \ln(p_{10}+p_{00}) - \ln(p_{01}+p_{00}))k) \\ &= \sum_{i=0}^{+\infty} \frac{(2k)^i\big[\ln p_{00} - \ln(p_{10}+p_{00}) - \ln(p_{01}+p_{00})\big]^i}{i!}. \end{aligned}$$

Case 1: $p_{00}p_{11} - p_{10}p_{01} \geq 0$.

Since x^i is a strictly convex function for all $i > 1$, $x \in (0, \infty)$, and $\ln 1/p_{00} + \ln 1 - \ln(1+p_{10}/p_{00}) - \ln(1+p_{01}/p_{00}) \geq 0$, $\ln^i 1/p_{00} + \ln^i 1 - \ln^i(1+p_{10}/p_{00}) - \ln^i(1+p_{01}/p_{00}) > 0$. So $g_{ab}(k)/p_{00}^k \geq 0$ for $a, b \in \{0,1\}$. On the other hand, when $p_{00}p_{11} - p_{10}p_{01} \geq 0$, $\ln p_{00} - \ln(p_{10}+p_{00}) - \ln(p_{01}+p_{00}) \geq 0$. Hence, we can rewrite

$$\psi_1(k, \pi, \mathbf{p}) = \frac{1}{m^3 k^6} \sum_{i=0}^{+\infty} w_i k^i = \frac{1}{n^3 k^3} \sum_{i=0}^{+\infty} w_i k^i, \ \ w_i \geq 0.$$

When n is fixed, the second-order partial derivative of $\psi_1(k, \pi, \mathbf{p})$ with respect to k is $\partial^2 \psi_1(k)/\partial k^2 = (1/n^3)\sum_{i=0}^{+\infty} w_i(i-3)(i-4)k^{i-5}$. As m is fixed, $\partial^2 \psi_1(k)/\partial k^2 = (1/m^3)\sum_{i=0}^{+\infty} w_i(i-6)(i-7)k^{i-8}$. Obviously, $\partial^2 \psi_1(k, \pi, \mathbf{p})/\partial k^2 \geq 0$ since all $w_i \geq 0$. Therefore, $\psi_1(k, \pi, \mathbf{p})$ is a convex function of k. Thus, $\psi_1(k, \pi, \mathbf{p})$ is a strictly convex function since $w_i > 0$ as $i > 3$.

Case 2: $p_{00}p_{11} - p_{10}p_{01} < 0$.

Rewrite $g_{ab}(k)/p_{00}^k$ as

$$\frac{g_{ab}(k)}{p_{00}^k} = \pi_{00}^{ab} + \sum_{i=1}^{\infty} \Big\{\pi_{10}^{ab} \ln^i\Big(1+\frac{p_{10}}{p_{00}}\Big) + \pi_{01}^{ab} \ln^i\Big(1+\frac{p_{01}}{p_{00}}\Big) + \pi_{11}^{ab}\Big[\ln^i\Big(\frac{1}{p_{00}}\Big) - \ln^i\Big(1+\frac{p_{10}}{p_{00}}\Big) - \ln^i\Big(1+\frac{p_{01}}{p_{00}}\Big)\Big]\Big\}\frac{k^i}{i!}.$$

Consider that $(a,b) \neq (1,1)$ and write $h_i(x) = \ln^i(x)$. For $i \geq 2$ and $x \in (1,2)$, the second-order derivative of $h_i(x)$ is

$$h_i^{''}(x) = \frac{i}{x^2}\ln^{i-2}(x)[i-1-\ln(x)] \geq 0,$$

and $h_i(x)$ is a convex function. Since $1/p_{00}+1 > 1+p_{10}/p_{00}+1+p_{01}/p_{00}$, we have

$$h_i(\frac{1}{p_{00}}) + h_i(1) > h_i(1+\frac{p_{10}}{p_{00}}) + h_i(1+\frac{p_{01}}{p_{00}}),$$

which is $\ln^i(1/p_{00}) - \ln^i(1+p_{10}/p_{00}) - \ln^i(1+p_{01}/p_{00}) > 0$. Therefore, the coefficient of k^i is positive for $i \geq 2$. For $i=1$, the coefficient of k^i is

$$\pi_{10}^{ab}\ln\left(1+\frac{p_{10}}{p_{00}}\right)+\pi_{01}^{ab}\ln\left(1+\frac{p_{01}}{p_{00}}\right)+\pi_{11}^{ab}\left[\ln\left(\frac{1}{p_{00}}\right)-\ln\left(1+\frac{p_{10}}{p_{00}}\right)-\ln\left(1+\frac{p_{01}}{p_{00}}\right)\right].$$

When $(a,b) \neq (1,1)$, we have $\pi_{10}^{ab}+\pi_{01}^{ab} > \pi_{11}^{ab}$ and verify that the coefficient of k^i is positive for all $i \geq 1$.

If

$$S_1(k) = \frac{1}{m^3k^6}\frac{g_{00}}{p_{00}^k}\frac{g_{10}}{p_{00}^k}\frac{g_{01}}{p_{00}^k} \quad \text{and} \quad S_2(k) = \frac{p_{00}^{2k}}{(p_{00}+p_{10})^{2k}(p_{00}+p_{01})^{2k}}\frac{g_{11}}{p_{00}^k},$$

then $\psi_1(k,\pi,\mathbf{p}) = S_1(k)S_2(k)$. Following the proof in Case 1, $S_1(k)$ is a convex function. We proceed to verify that $S_2(k)$ is also a positive convex function.

As is derived above,

$$\frac{g_{11}}{p_{00}^k} = \pi_{00}^{11} + \sum_{i=1}^{\infty}\left\{\pi_{10}^{11}\ln^i\left(1+\frac{p_{10}}{p_{00}}\right)+\pi_{01}^{11}\ln^i\left(1+\frac{p_{01}}{p_{00}}\right)\right.$$
$$\left.+\pi_{11}^{11}\left[\ln^i\left(\frac{1}{p_{00}}\right)-\ln^i\left(1+\frac{p_{10}}{p_{00}}\right)-\ln^i\left(1+\frac{p_{01}}{p_{00}}\right)\right]\right\}\frac{k^i}{i!}.$$

Write

$$\frac{g_{11}}{p_{00}^k} - \pi_{00}^{11} = \sum_{i=1}^{\infty}\alpha_i k^i,$$

the result of Case 1 gives $\alpha_i > 0$ for $i \geq 2$.

By the Condition, $\max\{\pi_{10}^{11},\pi_{01}^{11}\} < \pi_{11}^{11}$, so we have

$$\alpha_1 < [\max\{\pi_{10}^{11},\pi_{01}^{11}\} - \pi_{11}^{11}]\ln(\frac{(p_{00}+p_{10})(p_{00}+p_{01})}{p_{00}}) < 0.$$

For $k=1$, $g_{11}/p_{00} - \pi_{00}^{11}$ reduces to

$$\frac{g_{11}}{p_{00}} - \pi_{00}^{11} = \pi_{01}^{11}\frac{p_{01}}{p_{00}} + \pi_{10}^{11}\frac{p_{10}}{p_{00}} + \pi_{11}^{11}\frac{p_{11}}{p_{00}} > 0.$$

Then we obtain $\sum_{i=1}^{\infty}\alpha_i = g_{11}/p_{00}^k - \pi_{00}^{11} > 0$. Let $\phi = p_{00}^2/((p_{00}+p_{10})^2(p_{00} +p_{01})^2)$, and rewrite $S_2(k)$ as

$$\begin{aligned} S_2(k) &= \frac{p_{00}^{2k}}{(p_{00}+p_{10})^{2k}(p_{00}+p_{01})^{2k}}\frac{g_{11}}{p_{00}^k} \\ &= \phi^k\Big[\pi_{00}^{11} + \sum_{i=1}^{\infty}\alpha_i k^i\Big] = \pi_{00}^{11}\phi^k + \sum_{i=1}^{\infty}\alpha_i\phi^k k^i. \end{aligned}$$

The second-order derivative of $S_2(k)$ is

$$S_2^{''}(k)=\phi^k\Big[\pi_{00}^{11}\ln^2\phi+\alpha_1(k\ln^2\phi+2\ln\phi)+\sum_{i=2}^{\infty}\alpha_i k^{i-2}(k^2\ln^2\phi+2ik\ln\phi+i(i-1))\Big].$$

Since $\phi = p_{00}^2/[(p_{00}+p_{10})^2(p_{00}+p_{01})^2]$, we have $-\ln\phi = 2\ln(1+(p_{01}p_{10} -p_{00}p_{11})/p_{00})$. As we know, if $x\in[0,1)$, then $2\ln(1+x)\in(x,2x)$. Therefore,

$$\begin{aligned} -\ln\phi &\in (\frac{p_{01}p_{10}-p_{00}p_{11}}{p_{00}}, 2\frac{p_{01}p_{10}-p_{00}p_{11}}{p_{00}}), \\ \ln\phi &\in (-2\frac{p_{01}p_{10}-p_{00}p_{11}}{p_{00}}, -\frac{p_{01}p_{10}-p_{00}p_{11}}{p_{00}}). \end{aligned}$$

Since group testing is usually used for rare traits, it is reasonable to assume $\max\{p_{01},p_{10}\} < 0.05$ and $p_{00}\geq 0.9$, and then $\ln\phi > -0.00556$.

In group testing, it is reasonable to assume the maximum tolerate group size $k_{\max} < 100$, so $k\ln^2\phi + 2\ln\phi = \ln\phi(k\ln\phi+2) < 0$ and $k^2\ln^2\phi + 2ik\ln\phi + i(i-1)) > 0, \forall i\geq 2$. Then we have $\alpha_1(k\ln^2\phi+2\ln\phi) > 0$. The second-order derivative of $S_2(k)$ is

$$\begin{aligned} S_2^{''}(k) = \phi^k\Big[\pi_{00}^{11}\ln^2\phi+\alpha_1(k\ln^2\phi+2\ln\phi)+\sum_{i=2}^{\infty}\alpha_i k^{i-2}(k^2\ln^2\phi+2ik\ln\phi+i(i-1))\Big] \\ > 0. \end{aligned}$$

Therefore, $S_2(k)$ is a convex function. If $k_{i,opt} = \text{argmin}_k S_i(k)$, since $\psi_1(k,\pi,\mathbf{p}) = S_1(k)S_2(k)$, the optimum group size belongs to the interval $[\min\{k_{1,opt},k_{2,opt}\}, \max\{k_{1,opt},k_{2,opt}\}]$. The proof is completed.

Appendix D. Proof of Theorem 2 and 3

Since

$$\begin{aligned} \psi(k,\pi,\mathbf{p}) \propto \psi_1(k,\pi,\mathbf{p}) &= \frac{g_{00}g_{01}g_{10}g_{11}}{m^3k^6p_{00}^{2k}(p_{00}+p_{10})^{2k}(p_{00}+p_{01})^{2k}} \\ &= \frac{p_{00}^{2k}}{m^3k^6(p_{00}+p_{10})^{2k}(p_{00}+p_{01})^{2k}}\frac{g_{00}}{p_{00}^k}\frac{g_{10}}{p_{00}^k}\frac{g_{01}}{p_{00}^k}\frac{g_{11}}{p_{00}^k}, \end{aligned}$$

$\psi(1,\pi,\mathbf{p}) > \psi(2,\pi,\mathbf{p})$ is equivalent to $\psi_1(1,\pi,\mathbf{p}) > \psi_1(2,\pi,\mathbf{p})$.

Using the notation in the main text, for any given $\delta \in (\max\{p_{10}/p_{00} + p_{11}/p_{00}, p_{01}/p_{00} + p_{11}/p_{00}\}, 1)$ and $a, b \in \{0,1\}$, let $h_{ab}(\lambda) = \xi_{ab}\delta^2 + \eta_{ab}(2-\lambda)\delta - \zeta_{ab}(\lambda - 1)$. Then $h_{ab}(\lambda) = 0$ has a solution in λ. Denote it by λ_{ab} where $\lambda_{ab} = (\xi_{ab}\delta^2 + 2\eta_{ab}\delta + \zeta_{ab})/(\eta_{ab}\delta + \zeta_{ab})$, $a, b \in \{0,1\}$. Because $h_{ab}(\lambda)$ is a strictly decreasing function of λ and $h_{ab}(1) > 0$, $\lambda_{ab} > 1$.

Let

$$f_1(\delta) = \frac{p_{00}^2}{2^6(p_{00}+p_{10})^2(p_{00}+p_{01})^2} \prod_{a,b\in\{0,1\}} \lambda_{ab}(\delta)$$

and

$$f_2(\delta) = \frac{p_{00}^2}{2^3(p_{00}+p_{10})^2(p_{00}+p_{01})^2} \prod_{a,b\in\{0,1\}} \lambda_{ab}(\delta).$$

Since $\lambda_{ab}(\delta) > 0$ and

$$\frac{\partial \lambda_{ab}(\delta)}{\partial \delta} = \frac{2\xi_{ab}\delta + 2\eta_{ab}}{\eta_{ab}\delta + \zeta_{ab}} - \eta_{ab}\frac{\xi_{ab}\delta^2 + 2\eta_{ab}\delta + \zeta_{ab}}{(\eta_{ab}\delta + \zeta_{ab})^2} = \frac{\xi_{ab}\eta_{ab}\delta^2 + 2\xi_{ab}\zeta_{ab}\delta + \eta_{ab}\zeta_{ab}}{(\eta_{ab}\delta + \zeta_{ab})^2} > 0,$$

$f_j(\delta)$ is an increasing function of δ, and goes to $p_{00}^2/(2^6(p_{00}+p_{10})^2(p_{00}+p_{01})^2)$ as $\delta \to 0^+$, to $+\infty$ as $\delta \to +\infty$, $j = 1, 2$. The equation $f_j(\delta) = 1$ has a unique solution for δ, denote it by δ_j, $j = 1, 2$.

After some algebras, we have

$$\begin{aligned}
\frac{g_{ab}|_{k=2}}{p_{00}^2} &= \pi_{00}^{ab} + \pi_{10}^{ab}\Big[\Big(1 + \frac{p_{10}}{p_{00}}\Big)^2 - 1\Big] + \pi_{01}^{ab}\Big[\Big(1 + \frac{p_{01}}{p_{00}}\Big)^2 - 1\Big] \\
&\quad + \pi_{11}^{ab}\Big[1 + \Big(1 + \frac{p_{01}}{p_{00}} + \frac{p_{10}}{p_{00}} + \frac{p_{11}}{p_{00}}\Big)^2 - \Big(1 + \frac{p_{10}}{p_{00}}\Big)^2 - \Big(1 + \frac{p_{01}}{p_{00}}\Big)^2\Big] \\
&= \pi_{00}^{ab} + 2\pi_{10}^{ab}\frac{p_{10}}{p_{00}} + 2\pi_{01}^{ab}\frac{p_{01}}{p_{00}} + 2\pi_{11}^{ab}\frac{p_{11}}{p_{00}} + \pi_{10}^{ab}(\frac{p_{10}}{p_{00}})^2 + \pi_{01}^{ab}(\frac{p_{01}}{p_{00}})^2 \\
&\quad + \pi_{11}^{ab}\Big[2\frac{p_{01}+p_{11}}{p_{00}}\frac{p_{10}+p_{11}}{p_{00}} - (\frac{p_{11}}{p_{00}})^2\Big], \\
\frac{g_{ab}|_{k=2}}{p_{00}^2} &- \lambda_{ab}\frac{g_{ab}|_{k=1}}{p_{00}} \\
&= (2 - \lambda_{ab})\left[\pi_{10}^{ab}\frac{p_{10}}{p_{00}} + \pi_{01}^{ab}\frac{p_{01}}{p_{00}} + \pi_{11}^{ab}\frac{p_{11}}{p_{00}}\right] + \pi_{10}^{ab}(\frac{p_{10}}{p_{00}})^2 + \pi_{01}^{ab}(\frac{p_{01}}{p_{00}})^2 \\
&\quad + \pi_{11}^{ab}\left[2\frac{p_{01}+p_{11}}{p_{00}}\frac{p_{10}+p_{11}}{p_{00}} - (\frac{p_{11}}{p_{00}})^2\right] - (\lambda_{ab} - 1)\pi_{00}^{ab}.
\end{aligned}$$

Then,

$$\frac{g_{00}|_{k=2}}{p_{00}^2} - \lambda_{00}\frac{g_{00}|_{k=1}}{p_{00}} < \xi_{00}\delta^2 + \eta_{00}\delta(2-\lambda_{00}) - \zeta_{00}(\lambda_{00}-1) = 0.$$
$$\frac{g_{10}|_{k=2}}{p_{00}^2} - \lambda_{10}\frac{g_{10}|_{k=1}}{p_{00}} < \xi_{10}\delta^2 + \eta_{10}\delta(2-\lambda_{10}) - \zeta_{10}(\lambda_{10}-1) = 0.$$
$$\frac{g_{01}|_{k=2}}{p_{00}^2} - \lambda_{01}\frac{g_{01}|_{k=1}}{p_{00}} < \xi_{01}\delta^2 + \eta_{01}\delta(2-\lambda_{01}) - \zeta_{01}(\lambda_{01}-1) = 0.$$
$$\frac{g_{11}|_{k=2}}{p_{00}^2} - \lambda_{11}\frac{g_{11}|_{k=1}}{p_{00}} < \xi_{11}\delta^2 + \eta_{11}\delta(2-\lambda_{11}) - \zeta_{11}(\lambda_{11}-1) = 0.$$

So, $g_{ab}|_{k=2}/p_{00}^2 < \lambda_{ab}g_{ab}|_{k=1}/p_{00}$, $a, b \in \{0,1\}$.

When m is fixed,

$$\psi_1(2,\pi,\mathbf{p}) < \psi_1(1,\pi,\mathbf{p})\frac{p_{00}^2}{2^6(p_{00}+p_{10})^2(p_{00}+p_{01})^2}\prod_{a,b\in\{0,1\}}\lambda_{ab}(\delta_1) = \psi_1(1,\pi,\mathbf{p}).$$

When n is fixed,

$$\psi_1(2,\pi,\mathbf{p}) < \psi_1(1,\pi,\mathbf{p})\frac{p_{00}^2}{2^3(p_{00}+p_{10})^2(p_{00}+p_{01})^2}\prod_{a,b\in\{0,1\}}\lambda_{ab}(\delta_2) = \psi_1(1,\pi,\mathbf{p}).$$

This completes the proofs.

References

Barcellos, L. F., Klitz, W., Field, L. L., Tobias, R., Bowcock, A. M., Wilson, R., Nelson, M. P., Nagatomi, J. and Thomson, G. (1997). Association mapping of disease loci, by use of a pooled DNA genomic screen. *Amer. J. Hum. Genet.* **61**, 734-747.

Bilder, C., Tebbs, J. and Chen, P. (2010). Informative retesting. *J. Amer. Statist. Assoc.* **105**, 942-955.

Brookmeyer, R. (1999). Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics* **55**, 608-612.

Cahoon-Young, B., Chandler, A., Livermore, T., Gaudino, J. and Benjamin, R. (1989). Sensitivity and specificity of pooled versus individual sera in a human immunodeficiency virus antibody prevalence study. *J. Clin. Microbiol.* **27**, 1893-1895.

Chen, C. L. and Swallow, W. H. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics* **46**, 1035-1046.

Chien, C. C., Huang, S. F. and Lung, F. W. (2009). Maximally efficient two-stage screening: Determining intellectual disability in Taiwanese military conscripts. *J. Multidiscip. Healthc.* **2**, 39-44.

Datta, S. D., Sternberg, M., Hohnson, R. E., Berman, S., Papp, J. R., Mcquilla, G. and Weinstock, H. (2007). Gonorrhea and chlamydia in the United States among persons 14 to 39 years of age, 1999 to 2002. *Ann. Internal Medicine* **147**, 89-96.

Delaigle, A. and Meister, A. (2011). Nonparametric regression analysis for group testing data. *J. Amer. Statist. Assoc.* **106**, 640-650.

Dorfman, R. (1943). The detection of defective members of large populations. *Ann. Math. Statist.* **14**, 436-440.

Fahey, J., Ourisson, P. and Degnan, F. (2006). Pathogen detection, testing and control in fresh broccoli sprouts. *Nutr. J.* **5**, 13.

Graff, L. E. and Roeloffs, R. (1972). Group testing in the presence of test errors: An extension of the Dorfman procedure. *Technometrics* **14**, 113-122.

Gastwirth, J. and Johnson, W. (1994). Screening with cost-effective quality control: Potential applications to HIV and drug testing. *J. Amer. Statist. Assoc.* **89**, 972-981.

Gu, W, Lampman, R. and Novak, R. (2004). Assessment of arbovirus vector infection rates using variable size pooling. *Med. Vet. Entomol.* **18**, 200-204.

Hepworth, G. and Watson, R. (2009). Debiased estimation of proportions in group testing. *J. Roy. Statist. Soc. Ser. C* **58**, 105-121.

Hung, M. and Swallow, W. H. (1999). Robustness of group testing in the estimation of proportions. *Biometrics* **55**, 231-237.

Hughes-Oliver, J. M. and Rosenberger, W. F. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika* **87**, 315-327.

Hughes-Oliver, J. M. and Swallow, W. H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *J. Amer. Statist. Assoc.* **89**, 982-993.

Jarvis, L. M., Dow, B. C., Cleland, A., Davidson, F., Lycett, C., Morris, K., Webb, B., Jordan, A. and Petrik, J. (2005). Detection of HCV and HIV-1 antibody negative infections in Scottish and Northern Ireland blood donations by nucleic acid amplification testing. *Vox Sang.* **89**, 128-134.

Kainkaryam, R. and Woolf, P. (2009). Pooling in high-throughput drug screening. *Curr. Opin. Drug. Disc.* **12**, 339-350.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Statist.* **2**, 849-879.

Litvak, E., Tu, X. M. and Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *J. Amer. Statist. Assoc.* **89**, 424-434.

Liu, A., Liu, C., Zhang, Z. and Albert, P. (2012). Optimality of group testing in the presence of misclassification. *Biometrika* **99**, 254-251.

McMahan, C., Tebbs, J. and Bilder, C. (2012). Two-dimensional informative array testing. *Biometrics* **68**, 793-804.

Pritchard, N. and Tebbs, J. (2011). Estimating disease prevalence using inverse binomial pooled testing. *J. Agricultural, Biological, and Environmental Statistics* **16**, 70-87.

Ridout, M. S. (1995). Three-stage designs for seed testing experiments. *J. Roy. Statist. Soc. Ser. C* **44**, 153-162.

Sobel, M. and Elashoff, R. (1975). Group testing with a new goal, estimation. *Biometrika* **62**, 181-193.

Tebbs, J. M., McMahan, C. S. and Bilder, C. R. (2013). Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project. *Biometrics* **69**, 1064-1073.

Tu, X. M., Litvak, E. and Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika* **82**, 287-289.

Xie, M., Tatsuoka, K., Sacks, J. and Yound, S. S. (2001). Group testing with blockers and synergism. *J. Amer. Statist. Assoc.* **96**, 92-102.

Key Lab of Systems Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

E-mail: liqz@amss.ac.cn

Biostatistics and Bioinformatics Branch, NICHD, NIH, Bethesda, Maryland, U.S.A.

E-mail: liua@mail.nih.gov

School of Mathematics and Statistics, Guangxi Normal University, Guilin 541004, China.

E-mail: wjxiong@mailbox.gxnu.edu.cn

(Received May 2015; accepted April 2016)