Published in final edited form as:

Stat Med. 2009 October 15; 28(23): 2912–2928. doi:10.1002/sim.3678.

Global goodness-of-fit tests for group testing regression models

Peng Chen¹, Joshua M. Tebbs^{2,*†}, and Christopher R. Bilder^{3,‡}

¹ Takeda Global Research & Development Center, Inc., 675 North Field Drive, Lake Forest, IL 60045, U.S.A

² Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A

³ Department of Statistics, University of Nebraska, Lincoln, NE 68583, U.S.A

SUMMARY

In a variety of biomedical applications, particularly those involving screening for infectious diseases, testing individuals (e.g., blood/urine samples, etc.) in pools has become a standard method of data collection. This experimental design, known as group testing (or pooled testing), can provide a large reduction in testing costs and can offer nearly the same precision as individual testing. To account for covariate information on individual subjects, regression models for group testing data have been proposed recently. However, there are currently no tools available to check the adequacy of these models. In this paper, we present various global goodness-of-fit tests for regression models with group testing data. We use simulation to examine small-sample size and power properties of the tests for different pool composition strategies. We illustrate our methods using two infectious disease data sets, one from an HIV study in Kenya and one from the Infertility Prevention Project.

Keywords

Infertility Prevention Project; IOS test; latent binary response; logistic regression; pooled testing; score test

1. INTRODUCTION

When screening subjects for sexually transmitted diseases such as HIV, it is often more practical, at least initially, to test subjects in pools rather than to test them individually. Whether the goal is to estimate the prevalence of infection or to identify those who are infected, the argument for testing subjects in pools through group testing has been made repeatedly in the statistical literature [1]–[5], and the infectious disease literature is replete with empirical evidence showing the profound advantages of pooling; e.g., [6]–[9]. The advantages of group testing for estimation and identification in drug discovery are also well known [10,11].

In this paper, we consider regression models for a pooled binary response from group testing when covariate information is available for each subject in the pool. This type of model falls outside the usual binary regression model framework because the individual binary responses (e.g., infection statuses, etc.) are not necessarily observed. Motivated by a Kenyan HIV study [6], Vansteelandt et al. [1] propose a likelihood-based approach to model the

*Correspondence to: Joshua M. Tebbs, Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A.

†E-mail: tebbs@stat.sc.edu

‡Url: www.chrisbilder.com

probability of infection, as a function of fixed subject-specific covariates, using the observed initial pooled testing results. Xie [2] presents a conceptually different approach to fit fixed effects group testing regression models by treating individual statuses as latent and using the EM algorithm. This approach is more flexible than that in [1] because, among other things, one can incorporate testing results from retests; e.g., additional results obtained from retesting subsets of positive pools. More recently, Chen et al. [12] propose a generalization of the approach in [1] to include effects which are best regarded as random.

Modeling group testing data has been an important advance, because it allows one to take advantage of risk factors through the covariates while acknowledging population heterogeneity. This sharply contrasts with previous research in group testing which has largely required the individual probability of positivity to be the same for each subject. However, despite this methodological advance, there are currently no analytical tools one can use to assess the fit of the model. With this in mind, we propose global goodness-of-fit (GOF) tests tailored for use with the fixed effects group testing regression model proposed by Vansteelandt et al. [1]. These tests can provide needed guidance to assess the fit of a model, because graphical assessments of adequacy for this type of model are difficult to make. Our overarching approach in this paper is to consider analytical procedures which test whether or not a particular model is appropriate by detecting general departures from it.

Subsequent sections of this paper are organized as follows. In Section 2, we review the group testing regression modeling approach of Vansteelandt et al. [1]. In Section 3, we discuss the components of model adequacy and present four global GOF tests for group testing regression models. In Section 4, we consider different pool composition strategies and use simulation to characterize the performance of the tests. In Section 5, we apply our GOF procedures to two infectious disease data sets, one from the Kenyan HIV study in [1] and one from the Infertility Prevention Project. In Section 6, we summarize our findings, describe other approaches, and discuss future research in this area. Additional mathematical details are catalogued in the appendices.

2. NOTATION AND ASSUMPTIONS

Suppose that N individuals (e.g., blood/urine specimens, etc.) are drawn from a large population and that each individual is assigned to exactly one of n pools. Denote by c_i the pool size for the i th pool and let $Y_{ij} = 1$ if the j th individual in the i th pool is positive, $Y_{ij} = 0$ otherwise, for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, c_i$, so that $N = \sum_{i=1}^n c_i$. We assume throughout that the individual statuses Y_{ij} are independent and let $\mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2}, \dots, x_{ijp})'$ denote the $(p+1) \times 1$ fixed covariate vector for the j th individual in the i th pool. For the model formulation in [1], an arbitrary link function between the mean of Y_{ij} and the covariate $\mathbf{x}_{ij}$ is allowed. We restrict attention to the logistic link function, although it is possible to extend our work to handle other link functions. Specifically, we write

$$p_{ij} \equiv p_{ij}(\beta, \mathbf{x}_{ij}) = E(Y_{ij} | \mathbf{x}_{ij}; \beta) = \frac{\exp(\beta' \mathbf{x}_{ij})}{1 + \exp(\beta' \mathbf{x}_{ij})}, \quad (1)$$

where β is the $(p+1) \times 1$ fixed effects regression parameter that is to be estimated from the pooled testing results. For $i = 1, 2, \dots, n$, define $Z_i = 1$, if the testing result for the i th pool indicates the presence of at least one positive individual in the pool and $Z_i = 0$ otherwise. Similar to Vansteelandt et al. [1], we treat Z_i as an observable whereas the individual statuses Y_{ij} are not observed. We assume that the sensitivity (γ_1) and the specificity (γ_2) of

the diagnostic test (e.g., ELISA test, etc.) used to make pooled determinations are constants close to 1 and that they do not depend on c_i so that

$$p_i(\beta) = E(Z_i | \mathbf{x}_i; \beta) = \gamma_1 + \gamma_{12} \prod_{j=1}^{c_i} \{1 - p_{ij}(\beta, \mathbf{x}_{ij})\}, \quad (2)$$

where $\mathbf{x}_i = (\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \dots, \mathbf{x}'_{ic_i})'$ and $\gamma_{12} = 1 - \gamma_1 - \gamma_2$. Because the latent statuses Y_{ij} are independent, the log-likelihood function of β , based on the observed data $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)'$, can be written as

$$l(\beta | \mathbf{Z}) = \sum_{i=1}^n \left[Z_i \log \left\{ \frac{p_i(\beta)}{1 - p_i(\beta)} \right\} + \log \{1 - p_i(\beta)\} \right]. \quad (3)$$

One can maximize the log-likelihood in (3) to obtain the maximum likelihood estimate (MLE) for β , denoted throughout by $\hat{\beta}$. In practice, $\hat{\beta}$ can be computed using the quasi-Newton method described in [1], which we implement using the R function `optim`.

3. GLOBAL GOODNESS-OF-FIT TESTS

Because the latent individual statuses Y_{ij} are independent, the group testing regression model in (2) is literally “induced” by the logistic regression model in (1), so we focus on assessing the validity of (1). The essential components of fit for the logistic model can be specified by the following assumptions [13]:

A1. The logit transformation is the correct function linking the covariates with the conditional mean.

A2. The linear predictor $\beta' \mathbf{x}_{ij}$ is correct; that is, one does not need to include additional covariates, transformations of variables, interactions, etc.

A3. The variance of Y_{ij} , conditional on $\mathbf{x}_{ij}$, is $\text{var}(Y_{ij} | \mathbf{x}_{ij}) = p_{ij}(\beta, \mathbf{x}_{ij}) \{1 - p_{ij}(\beta, \mathbf{x}_{ij})\}$.

Instead of checking A1–A3 individually (assumptions which are obviously confounded), we focus on the assessment of overall GOF with pooled testing response; that is, we aim to detect violations of any of the assumptions listed above. In doing so, we consider only the situation wherein there is no “natural grouping” of the individuals. This is likely the case when at least one of the covariates is continuous, so that the number of covariate patterns is close to or equal to N , the number of individuals.

We now describe four global GOF tests for (1) with pooled testing response. As the individual statuses Y_{ij} are not observed, GOF is assessed using the available pooled testing results Z_i . Other testing approaches are discussed in Section 6.

3.1. Pearson χ^2 test

The first GOF test statistic we consider is a Pearson χ^2 -type statistic, defined by

$$G = G(\hat{\beta}) = \sum_{i=1}^n \frac{\{Z_i - p_i(\hat{\beta})\}^2}{p_i(\hat{\beta})\{1 - p_i(\hat{\beta})\}}, \quad (4)$$

where $p_i(\hat{\beta})$ is an estimate of the probability that the i th pool tests positive, based on the assumed model. As with its individual data analogue, (4) can be viewed as the sum of squared Pearson residuals and is a measure of deviation of the observed data from the assumed model. For logistic regression models with individual data; i.e., when $c_i = 1$, Hosmer et al. [13] show via simulation that the Pearson χ^2 statistic, using a normal approximation, enjoys good size and power properties when compared to other commonly used GOF statistics for logistic regression. Also using individual data, Osius and Rojek [14] derive the asymptotic normal distribution of the Pearson χ^2 statistic in very general settings.

One could mimic the approach of [14] and establish the asymptotic normality of our statistic defined in (4). However, we instead use the scaled chi-square distribution $b\chi_\nu^2$ as a reference, rather than a normal distribution, to achieve better distributional properties with finite samples [13,15]. The $b\chi_\nu^2$ distribution is commonly used to approximate the distribution of nonnegative random variables [16]. The values $b > 0$ and $\nu > 0$ are estimated by equating the first two (approximate) moments of G and those of the scaled χ^2 distribution. The approach we use to obtain the moments of G follows the same spirit as the approach in [14], although the calculations are more complicated when considering group testing responses. It is understood that asymptotic properties apply when the number of pools n is large.

We sketch the derivation of the first two moments of $G = G(\hat{\beta})$ here and provide full details in Appendix A. Suppose that the true value β_0 is not on the boundary of the parameter space and denote the score function by $S(\beta) = \partial(\ell(\beta)|Z)/\partial\beta$. Expanding $G(\hat{\beta})$ about β_0 , we have

$$G(\hat{\beta}) \approx G(\beta_0) + \left\{ \frac{\partial G(\beta)}{\partial \beta'} \bigg|_{\beta=\beta_0} \right\} (\hat{\beta} - \beta_0). \quad (5)$$

Under mild regularity conditions, it can be shown that

$$\hat{\mu}_G \equiv E\{G(\hat{\beta})\} \approx E\{G(\beta_0)\} = \sum_{i=1}^n E \left[\frac{\{Z_i - p_i(\beta_0)\}^2}{p_i(\beta_0)\{1 - p_i(\beta_0)\}} \right] = n$$

and that $\text{var}\{G(\hat{\beta})\} \approx \text{var}\{G(\beta_0)\} - C(\beta_0)' \mathcal{Q}^{-1}(\beta_0) C(\beta_0)$, where $\mathcal{Q}(\beta_0) = E\{S(\beta_0)S(\beta_0)'\}$ is the Fisher information matrix and

$$C(\beta_0) = E \left\{ \frac{\partial G(\beta)}{\partial \beta'} \bigg|_{\beta=\beta_0} \right\}.$$

We provide closed-form expressions for $\text{var}\{G(\beta_0)\}$, $\mathcal{Q}(\beta_0)$, and $C(\beta_0)$ in Appendix A.

Denote by $\hat{\sigma}_G^2$ the estimate of $\text{var}\{G(\hat{\beta})\}$, obtained by substituting $\hat{\beta}$ for β_0 in the above expressions. The testing procedure we propose is outlined as follows:

1. Calculate the Pearson statistic $G = G(\hat{\beta})$ and the approximate variance $\hat{\sigma}_G^2$.
2. Estimate the scale parameter b and degrees of freedom ν for the scaled χ^2 distribution by equating the first two moments of $b\chi_\nu^2$ and G ; that is, set $\hat{\mu}_G = b\nu$ and $\hat{\sigma}_G^2 = 2b^2\nu$. Solving for b and ν , we get $b = \hat{\sigma}_G^2 / 2\hat{\mu}_G$ and $\nu = 2\hat{\mu}_G^2 / \hat{\sigma}_G^2$.

3. Because G/b follows an approximate χ^2 distribution with ν degrees of freedom, the p-value for the test of GOF is $p = \text{pr}(\chi_\nu^2 \geq G/b)$.

3.2. Score test

For individual testing data, Stukel [17] extends the logistic regression model by adding two additional parameters α_1 and α_2 , on the logit scale, where $\alpha_1(\alpha_2)$ controls the shape of the mean function for positive (negative) values of the linear predictor $\beta' \mathbf{x}_{ij}$. In logistic regression, positive values of $\beta' \mathbf{x}_{ij}$ correspond to probabilities greater than 0.5, while negative values correspond to probabilities less than 0.5. However, in the low prevalence settings where group testing is commonly used, most subjects have small positive probabilities (often much smaller than 0.5). We thus modify Stukel's [17] approach and introduce only one additional parameter. Specifically, we embed the logistic model in (1) into a larger class of models indexed by the $(p+2) \times 1$ parameter vector $\theta = (\beta', \alpha_0)'$. For notational convenience, we let $\eta_{ij} = \beta' \mathbf{x}_{ij}$ denote the linear predictor so that the extended model can be written as

$$p_{ij}(\theta) = p_{ij}(\theta, \mathbf{x}_{ij}) = E(Y_{ij} | \mathbf{x}_{ij}; \theta) = \frac{\exp\{h_{\alpha_0}(\eta_{ij})\}}{1 + \exp\{h_{\alpha_0}(\eta_{ij})\}}, \quad (6)$$

where $h_{\alpha_0}(\eta_{ij}) = \eta_{ij}$ when $\eta_{ij} > 0$. When $\eta_{ij} \leq 0$,

$$h_{\alpha_0}(\eta_{ij}) = \begin{cases} -\alpha_0^{-1} \{\exp(\alpha_0 |\eta_{ij}|) - 1\}, & \alpha_0 > 0 \\ \eta_{ij}, & \alpha_0 = 0 \\ \alpha_0^{-1} \{\log(1 - \alpha_0 |\eta_{ij}|)\}, & \alpha_0 < 0. \end{cases}$$

It is worth emphasizing that $p_{ij}(\theta)$ depends on β only through the linear predictor η_{ij} , whereas the extra parameter α_0 controls the shape of the lower tail of the mean function $p_{ij}(\theta)$. A positive value of α_0 produces a curve with a lighter lower tail than the logistic curve, whereas a negative value of α_0 corresponds to a heavier lower tail. When $\alpha_0 = 0$, it is easy to see that the model in (6) reduces to the logistic model in (1). The adequacy of the logistic model in (1) can thus be assessed by testing $H_0: \alpha_0 = 0$.

Similar to Stukel [17], we formulate a score statistic to test H_0 for the group testing regression model in [1]. Specifically, we rewrite the mean of Z_i , $i = 1, 2, \dots, n$, as

$$p_i(\theta) = E(Z_i | \theta) = \gamma_1 + \gamma_{12} \prod_{j=1}^{c_i} \{1 - p_{ij}(\theta)\}, \quad (7)$$

so that the log-likelihood based on the extended model is

$$l(\theta | \mathbf{Z}) = l(\beta, \alpha_0 | \mathbf{Z}) = \sum_{i=1}^n \left[Z_i \log \left\{ \frac{p_i(\theta)}{1 - p_i(\theta)} \right\} + \log \{1 - p_i(\theta)\} \right].$$

The score function $S(\theta)$ is given by

$$S(\theta) = -\gamma_{12} \sum_{i=1}^n \left[\frac{\{Z_i - p_i(\theta)\}}{p_i(\theta) \{1 - p_i(\theta)\}} \prod_{j=1}^{c_i} \{1 - p_{ij}(\theta)\} \sum_{j=1}^{c_i} \frac{\partial p_{ij}(\theta) / \partial \theta}{1 - p_{ij}(\theta)} \right].$$

The statistic we propose is $T_S = S(\hat{\theta})$, where $\hat{\theta} = (\hat{\beta}', 0)'$ and $\hat{\beta}$ is the MLE of β calculated under H_0 , as described in Section 2. It can be shown that the explicit form of T_S is

$$T_S = \frac{\gamma_{12}}{2} \sum_{i=1}^n \left[\left\{ \frac{Z_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)} \right\} \prod_{j=1}^{c_i} (1 - \hat{p}_{ij}) \sum_{j=1}^{c_i} \hat{p}_{ij} \hat{\eta}_{ij}^2 I(\hat{\eta}_{ij} < 0) \right],$$

where $\hat{\eta}_{ij} = \hat{\beta}' \mathbf{x}_{ij}$, $I(\cdot)$ is the indicator function, and $\hat{p}_{ij}$ and $\hat{p}_i$ are the model-predicted probabilities obtained by substituting $\theta = \hat{\theta}$ into (6) and (7), respectively. When the number of pools n is large, under H_0 , T_S is approximately normal with zero mean. To derive the variance of T_S , denoted by $\sigma_{T_S}^2$, we partition $S(\theta)$ and the Fisher information matrix $\mathcal{Q}(\theta) = E\{S(\theta)S(\theta)'\}$ according to $\theta = (\beta', \alpha_0)'$, i.e., we write

$$S(\theta) = \frac{\partial l(\theta|\mathbf{Z})}{\partial \theta} = \begin{pmatrix} \partial l(\theta|\mathbf{Z}) / \partial \beta \\ \partial l(\theta|\mathbf{Z}) / \partial \alpha_0 \end{pmatrix} = \begin{pmatrix} S(\beta) \\ S(\alpha_0) \end{pmatrix}$$

and

$$I(\theta) = \begin{pmatrix} \mathbf{I}_{\beta\beta} & \mathbf{I}_{\beta\alpha_0} \\ \mathbf{I}_{\alpha_0\beta} & I_{\alpha_0\alpha_0} \end{pmatrix} = \begin{pmatrix} E\{S(\beta)S(\beta)'\} & E\{S(\beta)S(\alpha_0)\} \\ E\{S(\alpha_0)S(\beta)'\} & E\{S(\alpha_0)^2\} \end{pmatrix},$$

so that $\sigma_{T_S}^2 = I_{\alpha_0\alpha_0} - \mathbf{I}_{\alpha_0\beta} \mathbf{I}_{\beta\beta}^{-1} \mathbf{I}_{\beta\alpha_0}$. To obtain an estimate of $\sigma_{T_S}^2$, we substitute $\hat{\theta} = (\hat{\beta}', 0)'$ for θ in $\mathcal{Q}(\theta)$. Denote the resulting matrix by $\mathcal{Q}(\hat{\theta})$ and the corresponding partitions by $\hat{\mathbf{I}}_{\beta\beta}$, $\hat{\mathbf{I}}_{\beta\alpha_0}$, $\hat{\mathbf{I}}_{\alpha_0\beta}$, and $\hat{I}_{\alpha_0\alpha_0}$, respectively, so that $\hat{\sigma}_{T_S}^2 = \hat{I}_{\alpha_0\alpha_0} - \hat{\mathbf{I}}_{\alpha_0\beta} \hat{\mathbf{I}}_{\beta\beta}^{-1} \hat{\mathbf{I}}_{\beta\alpha_0}$. A closed-form expression for $\mathcal{Q}(\hat{\theta})$ is provided in Appendix B. The score test for $H_0: \alpha_0 = 0$ is a two sided test, with both positive and negative large values indicating lack of fit of the assumed model in (1). For implementation, we compare $T_S / \hat{\sigma}_{T_S}$ to the $\mathcal{N}(0, 1)$ reference distribution.

3.3. Hosmer-Lemeshow test

In logistic regression with individual data, the distribution of the Pearson χ^2 statistic can be approximated by a χ^2 distribution when the expected number of observations for each covariate pattern is not small. Unfortunately, this requirement is violated when the number of covariate patterns is close to or equal to the number of observations, the situation in which we are interested. One way to address this problem is to “group” the observed individual data so that asymptotic arguments can be applied to the “grouped data” [18]–[20]. We extend this idea to applications involving pooled binary responses.

When compared to other available GOF tests for individual testing data, the Hosmer-Lemeshow (HL) test is known to have lower power in detecting certain departures from the logistic model [13,15]. However, the HL statistic remains widely used, so we reevaluate its

potential use and include it for comparison purposes with pooled response data. For clarity, when we henceforth refer to the HL test, we exclusively use the term “pool” to identify an amalgamate of individuals for group testing. We use the term “group” to represent a set of the observed pools.

The HL statistic is computed as follows. First, we fit the pooled testing regression model defined in (1) and (2), obtaining the MLE $\hat{\beta}$ and model-predicted pool probabilities $\hat{p}_i$, $i = 1, 2, \dots, n$. Second, we partition the pooled testing responses Z_i , $i = 1, 2, \dots, n$, into m subgroups $g_1, g_2, \dots, g_m$ according to the ranking of the n values of $\hat{p}_i$. We denote the number of observations in each group by s_k , $k = 1, 2, \dots, m$. The HL statistic based on the pooled responses is then given by

$$T_{HL} = \sum_{k=1}^m \frac{(\sum_{i \in g_k} Z_i - \sum_{i \in g_k} \hat{p}_i)^2}{(\sum_{i \in g_k} \hat{p}_i) (1 - \sum_{i \in g_k} \hat{p}_i / s_k)}.$$

In general, we choose the subgroup sizes s_k to be equal. If n is not a multiple of m , we let $g_1, g_2, \dots, g_{m-1}$ each consist of $[n/m]$ pooled responses and let g_m consist of the remaining $n - (m - 1)[n/m]$ pooled responses. Using extensive simulation, Hosmer and Lemeshow [18] demonstrate that when the number of distinct covariate patterns values is close to the sample size, the null distribution of the HL statistic can be well approximated by a χ^2 distribution with $m - 2$ degrees of freedom, provided that $[n/m]$ is not small. We use the same reference distribution to diagnose GOF with pooled testing responses.

3.4. IOS test

Presnell and Boos [21] propose a general test for misspecification in parametric models by comparing the maximized “in-sample” likelihood with the maximized “out-of-sample” likelihood. We now describe how their methodology can be applied to diagnose GOF for the group testing regression model in [1]. Let $f(Z_i; \beta)$ denote the probability mass function of Z_i , $i = 1, 2, \dots, n$, that is, $f(Z_i; \beta)$ is a Bernoulli mass function with mean equal to (2). For convenience, in our notation of $f(Z_i; \beta)$ we have suppressed the dependence on the covariate vectors $\mathbf{x}_{ij}$, $j = 1, 2, \dots, c_i$. Let $\hat{\beta}$ denote the MLE of β , as described in Section 2, and let $\hat{\beta}_{(i)}$ denote the MLE of β when the i th (pooled) observation is deleted from the sample $\mathbf{Z}$. The test statistic proposed by Presnell and Boos [21] is the logarithm of the “in-and-out-of-sample” (IOS) likelihood ratio

$$\text{IOS} = \log \left\{ \frac{\prod_{i=1}^n f(Z_i; \hat{\beta})}{\prod_{i=1}^n f(Z_i; \hat{\beta}_{(i)})} \right\} = \sum_{i=1}^n \{l(Z_i; \hat{\beta}) - l(Z_i; \hat{\beta}_{(i)})\},$$

where $l(Z_i; \beta) = \log f(Z_i; \beta)$. Computing IOS can be extremely cumbersome for large n , so the authors in [21] derive an approximate version of the IOS statistic. For $i = 1, 2, \dots, n$, let $\dot{l}(Z_i; \beta) = \partial l(Z_i; \beta) / \partial \beta$, and $\ddot{l}(Z_i; \beta) = \partial^2 l(Z_i; \beta) / \partial \beta \partial \beta'$ denote the gradient and Hessian of $l(Z_i; \beta)$, respectively. The approximate form of the IOS statistic is

$$\text{IOS}_A = \text{tr} \{ \widehat{I(\beta)}^{-1} \widehat{B(\beta)} \},$$

where $\widehat{I}(\widehat{\beta}) = n^{-1} \sum_{i=1}^n -\{\ddot{l}(Z_i; \widehat{\beta})\}$ is the average observed information matrix and $\widehat{B}(\widehat{\beta}) = n^{-1} \sum_{i=1}^n \dot{l}(Y_i; \widehat{\beta}) \dot{l}(Z_i; \widehat{\beta})'$ is the sample covariance of the score function. Note that IOS_A does not require the leave-one-out estimates, so it is far simpler to compute when n is large.

Presnell and Boos [21] show that IOS_A is approximately normal with mean equal to $\text{tr}[E\{\hat{l}(\beta_0)\}^{-1}E\{\hat{B}(\beta_0)\}]$, where β_0 is the probability limit of $\widehat{\beta}$. When the assumed model is correct, this mean equals the dimension of β . However, the authors show that the asymptotic variance of IOS_A is too difficult for “routine use” and note that the convergence to normality can be very slow. These findings were reinforced by Campanu and Presnell [22], who used simulation to investigate the use of IOS_A with logistic regression for individual testing data. For implementation, the authors in [21] recommend using a parametric bootstrap to approximate the null distribution of IOS_A ; i.e., the distribution of IOS_A under the assumed model. We adopt this recommendation in our consideration of IOS_A for GOF with group testing data.

4. SIMULATION EVIDENCE

4.1. Pooling strategies

We consider two strategies to form pools for group testing. The first strategy is random pooling; i.e., individuals are assigned to pools at random regardless of their covariate values in $\mathbf{x}_{ij}$. The second strategy we consider is to form the pools homogeneously; i.e., individuals are assigned to pools based on their values of $\mathbf{x}_{ij}$, with “similar” individuals being pooled together. Vansteelandt et al. [1] showed that a more precise estimator of β results when pools consist of $\mathbf{x}_{ij}$ that are as similar as possible. We refer to this as homogeneous pooling. Bilder and Tebbbs [23] discuss the practical aspects of using both pooling strategies in various group testing applications, including infectious disease screening.

4.2. Simulation models

We consider three types of model misspecification. For each of the simulation models we consider, we aim to emulate the low-prevalence settings where group testing is commonly used. To be specific, we choose parameter settings in each model so that the proportion of positive individuals is between 0.05 and 0.10. The first type of departure we consider is the omission of a quadratic term. We generate individual responses Y_{ij} from the model

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2, \quad (8)$$

where x_{ij} is a continuous covariate generated from a $\mathcal{U}(-3, 3)$ distribution. The parameters β_0 , β_1 , and β_2 are determined by equations $\text{pr}(Y_{ij} = 1 | x_{ij} = 3) = 0.40$, $\text{pr}(Y_{ij} = 1 | x_{ij} = 0) = 0.0293$ and $\text{pr}(Y_{ij} = 1 | x_{ij} = -3) = a$ with $a = 0.0014, 0.01, 0.03, 0.05, 0.1$. As a increases from 0.0014 to 0.1, the effect of the omitted quadratic term is progressively pronounced. The case $a = 0.0014$ corresponds to the null model with $\beta_2 = 0$.

The second type of departure we consider is the omission of an interaction term between a binary and a continuous covariate. The true model is

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij1} x_{ij2}, \quad (9)$$

where x_{ij1} is generated from a Bernoulli(0.4) distribution and the continuous x_{ij2} is generated from a $\mathcal{U}(-3, 3)$ distribution. We choose $\beta_3 = 0, 0.3, 0.5, 0.8$, and 1 so that the interaction term is progressively important. The null model corresponds to $\beta_3 = 0$. For a given choice of β_3 , the other parameters $\beta_0, \beta_1, \beta_2$ are determined by equations $\text{pr}(Y_{ij} = 1 | x_{ij1} = 0, x_{ij2} = 0) = 0.02$, $\text{pr}(Y_{ij} = 1 | x_{ij1} = 1, x_{ij2} = 3) = 0.4$, and $\text{pr}(Y_{ij} = 1 | x_{ij1} = 1, x_{ij2} = 0) = 0.1$.

The third type of departure we consider is a misspecification of the link function. We specify models with different link functions using the modified Stukel's generalized logistic model in (6). Individual samples are generated from

$$\text{logit}(p_{ij}) = h_{\alpha_0}(\beta x_{ij}), \quad (10)$$

where $h_{\alpha_0}(\cdot)$ is as defined in Section 3.2 and x_{ij} is a continuous covariate generated from a $\mathcal{U}(0, 6)$ distribution. We choose $(\alpha_0, \beta) = (0, -1.2), (-0.5, -3), (-1, -3)$, and $(0.5, -1.2)$ so that GOF can be assessed for misspecified links taking on a variety of shapes. Among these parameter settings, $(\alpha_0, \beta) = (0, -1.2)$ corresponds to the null case, $(\alpha_0, \beta) = (-0.5, -3)$ and $(-1, -3)$ represent link functions with a heavier lower tail than the logistic function, while $(\alpha_0, \beta) = (0.5, -1.2)$ corresponds to link function with a lighter lower tail.

4.3. Simulation description

For each model in Section 4.2, we first simulate $N = 1000$ individual observations $(Y_{ij}, \mathbf{x}_{ij})$ and create group testing samples $(Z_i, \mathbf{x}_i)$, $i = 1, 2, \dots, n$, based on the simulated individual observations. We then repeat this procedure 500 times; that is, our simulation results are based on $M = 500$ group testing data sets, each consisting of $N = 1000$ individuals. For simplicity, we assume a common pool size $c_i = c$, for $i = 1, 2, \dots, n$, with $c = 2, 5$, and 10, so that $n = 500, 200$ and 100, respectively. We also assume that $\gamma_1 = \gamma_2 = 1$ (perfect testing), because the results are similar for other values of γ_1 and γ_2 that are reasonably close to 1. Imperfect tests are considered in Section 5.

When we form the pools for group testing, we use both random pooling (RP) and homogeneous pooling (HP), as described in Section 4.1. Random pools are created by assigning individual observations to pools in the order in which they were generated; homogeneous pools are created by sorting the covariates first and assigning individuals to pools based on this covariate ordering. HP is straightforward to implement when the simulation model has only one continuous covariate. For the model in (9) with dichotomous x_{ij1} and continuous x_{ij2} , we first sort the individuals by x_{ij1} . We then sort by x_{ij2} , within the two levels of x_{ij1} , and construct quasi-homogeneous pools based on these ordered individual samples, following the recommendation in [1]. Bilder and Tebbs [23] discuss other approaches to form homogeneous pools. Once the group testing samples have been formed,

the binary pooled responses Z_i are determined by $Z_i = I(\sum_{j=1}^c Y_{ij} > 0)$.

For the HL test, we take $m = 10$ (i.e., the decile of risk statistic). For the IOS_A test, we calculate p-values using a parametric bootstrap with $B = 200$ resamples. That is, for each simulated group testing data set, we first fit the group testing regression model in (2). We then generate 200 bootstrap resamples, treating the MLE $\hat{\beta}$ as the true value and keeping intact the individual covariate structure in each pool. The IOS_A statistics are calculated based on the original simulated data set and the bootstrap samples, which we denote by IOS_A and $\text{IOS}_A^{(b)}$, for $b = 1, 2, \dots, 200$, respectively. The p-value is calculated by $\#\{\text{IOS}_A^{(b)} \geq \text{IOS}_A\} / 200$.

4.4. Simulation results

Tables 1, 2, and 3, contain estimated sizes and powers from our simulations under models (8), (9), and (10), respectively. First, note that all four GOF statistics G , T_S , T_{HL} , and IOS_A confer the correct size for the null situations; estimated nominal $\alpha = 0.05$ size tests are within the margin of Monte Carlo error for all simulations. Second, with few exceptions, tests for GOF under HP are more powerful than the same tests with RP. This is perhaps not unexpected, based on the findings of [1,23], but it is interesting to see how large the gain in power can be when comparing HP to RP. If logistically feasible, it appears to be well worth the additional burden to construct homogeneous pools. Note that Bilder and Tebbs [23] describe situations where HP may be difficult or impossible to implement. Finally, smaller pool sizes generally provide larger power; however, using larger pool sizes does not affect the power for HP as much as it does for RP, especially for the first and second types of misspecification.

Among the four GOF statistics, each has excellent power to detect the omission of a quadratic term (Table 1) when HP is used. The same can be said for detecting a missing interaction term (Table 2), except for IOS_A . None of the four tests do particularly well at detecting a misspecified link function (Table 3). Our findings are not incongruous with those in [13,22], where the performance of similar GOF tests is investigated for logistic regression with individual testing data. For the misspecified link function with $\alpha_0 = 0.5$ (Table 3), there are some convergence problems with HP when $c = 10$. This is largely due to the “complete separation” problem, as described in [23] for group testing regression models. In short, complete separation means that the covariate(s) perfectly discriminate(s) between the binary pooled responses, which creates problems when trying to maximize (3). For the $\alpha_0 = 0.5$, $c = 10$, HP setting in Table 3, the nonconvergence rate is about 30 percent, so we do not report the results. No other convergence problems were observed.

5. APPLICATIONS

5.1. Kenya data

Verstraeten et al. [6] and Vansteelandt et al. [1] discuss a public-health study investigating HIV prevalence in rural Kenya. In this study, blood samples were collected from 740 pregnant women at four locations, and risk covariates were measured on each individual, including age, marital status, parity, and education level. Samples were first tested by an enzyme-linked immunosorbent assay (ELISA) test and positives were confirmed by a rapid assay test. Individual samples with a positive ELISA test result and a negative rapid assay test result were tested again with an ELISA to classify the sample as positive or negative. This sequential testing procedure was also implemented on pools of samples created at a central laboratory. Verstraeten et al. [6] report a significant cost reduction from using group testing while producing very similar prevalence estimates to those obtained from individual testing. To illustrate their regression methodology, Vansteelandt et al. [1] use the Kenya data and relate the individual prevalence to age through the logistic model

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 \text{Age}_{ij}. \quad (11)$$

Our goal is to assess GOF for this model using pooled responses from the study.

After removing observations with missing values, as in [1], there are $N = 706$ individual observations available for analysis. We consider both “chronological” and “ordered” pool compositions. Chronological pools are constructed by the date of collection whereas ordered pools are constructed based on the values of the age covariate. As pointed out by the authors

in [1], these compositions likely emulate random and homogeneous pooling, respectively. For both pooling strategies, we use $n = 101$ pools; 100 of size $c = 7$ and 1 of size $c = 6$, as in [1]. The sequential testing strategy used to elicit the pooled responses Z_i provides highly sensitive and specific assessments; we use $\gamma_1 = 1$ and $\gamma_2 = 0.9997$, values used also in [1]. Table 4 displays the parameter estimates and standard errors for the model in (11) and the GOF results for the four tests in Section 3. For the HL test, we use $m = 10$; for the IOS_A test, we use $B = 1000$ bootstrap resamples. All four GOF tests reach the same conclusion for both pooling strategies; specifically, there is no evidence of model misspecification.

5.2. Infertility Prevention Project data

The state of Nebraska takes part in the nationwide Infertility Prevention Project (IPP) through its Sexually Transmitted Diseases and Infertility Control Program. At a number of clinic sites throughout the state, urine or swab specimens are collected on individuals and are then transported to the Nebraska Public Health Laboratory in Omaha for chlamydia and gonorrhea testing. Each year, about 30,000 tests are performed in the state; we use the individual testing data from the first quarter of 2006 for illustration purposes here. Specifically, the data set consists of chlamydia and gonorrhea infection statuses for 6,138 individual subjects as well as many potential risk covariates. Chen et al. [12] analyzes this same data set, treating the site effect as random. However, because we are interested in fixed effects models, we do not take this perspective and ignore the site effect. For each infection (chlamydia and gonorrhea, treated separately), we consider the fixed effects part of the model in [12] and relate the infection statuses Y_{ij} to four covariates via

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 \text{Age}_{ij} + \beta_2 \text{Gender}_{ij} + \beta_3 \text{Urethritis}_{ij} + \beta_4 \text{Symptoms}_{ij}. \quad (12)$$

Our goal is to assess GOF of this model for both infections using pooled responses.

For each infection, we construct artificial pooled testing responses Z_i , as in [12], by grouping individuals into pools of size $c = 10$ (except one pool of size 8) according to the order of the testing date, thereby creating a pseudo-randomized composition of the covariates. To incorporate the effects of imperfect testing, pooled responses are simulated assuming that assay sensitivity and specificity are $\gamma_1 = 0.98$ and $\gamma_2 = 0.95$, respectively. These values are reasonable for both infections based on the empirical findings in [8,9]. Table 5 gives the values of the four GOF statistics presented in Section 3 for both infections. For the HL test, we use $m = 10$; for the IOS_A test, we use $B = 1000$ bootstrap resamples, as before. For chlamydia, each GOF statistic except T_{HL} provides a similar small p-value, indicating some evidence of model misspecification in (12). For gonorrhea, however, the GOF results suggest little evidence of misspecification.

6. DISCUSSION

We have developed four global GOF tests for group testing regression models and have illustrated the usefulness of these tests using data from two infectious disease applications. Among the four tests, the Pearson χ^2 and modified score tests have the highest power, and we recommend them for use as a global check for model adequacy. Of course, due to the information loss from pooling, these tests may not have sufficient power to detect certain types of misspecification. This is especially true when random pooling is used. The computation time involved for all tests is not overly prohibitive, possibly with the exception of the IOS_A test. Programs for data analysis are available by request.

Other GOF tests could be formulated for this problem; in fact, the four tests presented in this paper are only a subset of those we investigated. We initially considered formulating an unweighted sum of squares statistic, similar to the $\hat{S}$ statistic proposed in [13] for individual testing data. However, the statistic, when suitably normalized under group testing, did not provide a test with the correct size, and the test did not have higher power than the best tests herein. We also investigated a collection of bootstrapped statistics obtained from the predicted individual latent responses and the smoothed residual statistics discussed in [13,15]. However, the bootstrap tests were not powerful, and there were technical difficulties in formulating the smoothed residuals tests with latent binary responses.

In this paper, we have taken a “global departure approach” to diagnose GOF with group testing models. It would be worthwhile to examine directed tests which target specific departures from A1–A3, although we leave this to future research. Future work could also include the generalization of our GOF tests to incorporate information from retesting subsets of positive pools or from other pooling algorithms [4]. Depending on the specific testing strategy used, some of the GOF tests discussed herein could be modified accordingly. However, the mathematical details could prove to be markedly more formidable. Detecting lack of fit in mixed effects group testing regression models [12] also remains as an open problem.

Acknowledgments

The authors would like to thank the Associate Editor and two anonymous referees for helpful comments on earlier versions of this article. We are grateful to Dr. Stijn Vansteelandt and his colleagues for providing us with the Kenya HIV data. We also thank Dr. Peter Iwen, Dr. Steven Hinrichs, and Philip Medina for their consultation on the Infertility Prevention Project. This research was supported by Grant R01 AI067373 from the National Institutes of Health.

Contract/grant sponsor: National Institutes of Health; contract/grant number: R01 AI067373

References

1. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000; 56:1126–1133.10.1111/j.0006-341X.2000.01126.x [PubMed: 11129470]
2. Xie M. Regression analysis of group testing samples. *Statistics in Medicine*. 2001; 20:1957–1969.10.1002/sim.817 [PubMed: 11427952]
3. Gastwirth J, Johnson W. Screening with cost effective quality control: Potential applications to HIV and drug testing. *Journal of the American Statistical Association*. 1994; 89:972–981.
4. Kim H, Hudgens M, Dreyfuss J, Westreich D, Pilcher C. Comparison of group testing algorithms for case identification in the presence of testing error. *Biometrics*. 2007; 63:1152–1163.10.1111/j.1541-0420.2007.00817.x [PubMed: 17501946]
5. Hung M, Swallow W. Use of binomial group testing in tests of hypotheses for classification or quantitative covariables. *Biometrics*. 2000; 56:204–212.10.1111/j.0006-341X.2000.00204.x [PubMed: 10783797]
6. Verstraeten T, Farah B, Duchateau L, Matu R. Pooling sera to reduce the cost of HIV surveillance: A feasibility study in a rural Kenyan district. *Tropical Medicine and International Health*. 1998; 3:747–750.10.1046/j.1365-3156.1998.00293.x [PubMed: 9754671]
7. Pilcher C, Fiscus S, Nguyen T, Foust E, Wolf L, Williams D, Ashby R, O’Dowd J, McPherson J, Stalzer B, Hightow L, Miller W, Eron J, Cohen M, Leone P. Detection of acute infections during HIV testing in North Carolina. *New England Journal of Medicine*. 2005; 352:1873–1883. [PubMed: 15872202]
8. Kacena K, Quinn S, Hartman S, Quinn T, Gaydos C. Pooling of urine samples for screening for *Neisseria gonorrhoeae* by ligase chain reaction: Accuracy and application. *Journal of Clinical Microbiology*. 1998; 36:3624–3628. [PubMed: 9817885]

9. Kacena K, Quinn S, Howell M, Madico G, Quinn T, Gaydos C. Pooling urine samples for ligase chain reaction screening for genital *Chlamydia trachomatis* infection in asymptomatic women. *Journal of Clinical Microbiology*. 1998; 36:481–485. [PubMed: 9466763]
10. Remlinger K, Hughes-Oliver J, Young S, Lam R. Statistical design of pools using optimal coverage and minimal collision. *Technometrics*. 2006; 48:133–143.10.1198/004017005000000481
11. Hughes-Oliver, J. Pooling experiments for blood screening and drug discovery. In: Dean, Angela; Lewis, Susan, editors. *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*. Springer; New York: 2006.
12. Chen P, Tebbs J, Bilder C. Group testing regression models with fixed and random effects. *Biometrics*. 2009 in press. 10.1111/j.1541-0420.2008.01183.x
13. Hosmer D, Hosmer T, le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*. 1997; 16:965–980. [PubMed: 9160492]
14. Osius G, Rojek D. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*. 1992; 87:1145–1152.
15. le Cessie S, Van Houwelingen J. A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*. 1991; 47:1267–1282.
16. Cox, R.; Hinkley, D. *Theoretical Statistics*. Chapman and Hall/CRC; Boca Raton: 1974.
17. Stukel T. Generalized logistic models. *Journal of the American Statistical Association*. 1988; 83:426–431.
18. Hosmer D, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics: Theory and Methods*. 1980; 9:1043–1069.
19. Lemeshow S, Hosmer D. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*. 1982; 115:92–106. [PubMed: 7055134]
20. Hosmer D, Lemeshow S, Klar J. Goodness-of-fit testing for the multiple logistic regression analysis when the estimated probabilities are small. *Biometrical Journal*. 1988; 30:911–924.
21. Presnell B, Boos D. The IOS test for model misspecification. *Journal of the American Statistical Association*. 2004; 99:216–227.10.1198/016214504000000214
22. Capanu M, Presnell B. Misspecification tests for binomial and beta-binomial models. *Statistics in Medicine*. 2008; 27:2536–2554.10.1002/sim.3049 [PubMed: 17914713]
23. Bilder C, Tebbs J. Bias, efficiency, and agreement in group-testing regression models. *Journal of Statistical Computation and Simulation*. 2009; 79:67–80.10.1080/00949650701608990 [PubMed: 20046956]

APPENDIX A

We provide additional details on the derivation of the first two approximate moments for G in Section 3.1. Suppose that the true value β_0 is not on the boundary of the parameter space and denote the score function by $S(\beta) = \partial(\beta|Z)/\partial\beta$. Expanding $S(\hat{\beta})$ about β_0 , we have

$$\mathbf{0} = S(\hat{\beta}) \approx S(\beta_0) - \left\{ -\frac{\partial S(\beta)}{\partial \beta'} \bigg|_{\beta=\beta_0} \right\} (\hat{\beta} - \beta_0), \quad (\text{A.1})$$

where $\mathbf{0}$ is a $(p+1) \times 1$ vector of zeros. Under suitable regularity conditions and from (5), it follows from the Law of Large Numbers that

$$\frac{\partial G(\beta)}{\partial \beta} \bigg|_{\beta=\beta_0} \approx E \left\{ \frac{\partial G(\beta)}{\partial \beta} \bigg|_{\beta=\beta_0} \right\} = C(\beta_0), \quad (\text{A.2})$$

for large n . Approximating the observed information matrix $-\partial S(\beta)/\partial \beta'|_{\beta=\beta_0}$ by the Fisher information matrix $\mathcal{Q}(\beta_0) = E\{S(\beta_0)S(\beta_0)'\}$, and combining (5), (A.1), and (A.2), we write

$$G(\widehat{\beta}) \approx G(\beta_0) + C(\beta_0)' I^{-1}(\beta_0) S(\beta_0). \quad (\text{A.3})$$

Our approximation of the first two moments of G is based on (A.3). Note that in (A.3), the only random quantities are $G(\beta_0)$ and $S(\beta_0)$, and all other terms are constants which depend on the true parameter β_0 . Let $f(\mathbf{z}; \beta)$ denote the density function for $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)'$. In what follows, we denote expectation with respect to the density function $f(\mathbf{z}; \beta)$ by $E_\beta(\cdot)$ and expectation with respect to $f(\mathbf{z}; \beta_0)$ by $E(\cdot)$. Since $E\{S(\beta_0)\} = \mathbf{0}$, we have

$$E\{G(\widehat{\beta})\} \approx E\{G(\beta_0)\} = \sum_{i=1}^n E \left[\frac{\{Z_i - p_i(\beta_0)\}^2}{p_i(\beta_0) \{1 - p_i(\beta_0)\}} \right] = n.$$

The fact that $E\{G(\beta_0)\}$ does not depend on β_0 implies $E_\beta\{G(\beta)\} = n$, for all β . Therefore,

$$\frac{\partial}{\partial \beta} E_\beta\{G(\beta)\} = \frac{\partial}{\partial \beta} \int_{\mathcal{R}^n} G(\beta) f(\mathbf{z}; \beta) d\mathbf{z} = \mathbf{0}. \quad (\text{A.4})$$

Interchanging integration and differentiation, the right-hand side of (A.4) can be written as

$$\begin{aligned} \frac{\partial}{\partial \beta} \int_{\mathcal{R}^n} G(\beta) f(\mathbf{z}; \beta) d\mathbf{z} &= \int_{\mathcal{R}^n} \frac{\partial G(\beta)}{\partial \beta} f(\mathbf{z}; \beta) d\mathbf{z} + \int_{\mathcal{R}^n} G(\beta) \frac{\partial f(\mathbf{z}; \beta)}{\partial \beta} d\mathbf{z} \\ &= E_\beta \left\{ \frac{\partial G(\beta)}{\partial \beta} \right\} + \int_{\mathcal{R}^n} G(\beta) \frac{\partial \log\{f(\mathbf{z}; \beta)\}}{\partial \beta} f(\mathbf{z}; \beta) d\mathbf{z} \\ &= E_\beta \left\{ \frac{\partial G(\beta)}{\partial \beta} \right\} + E_\beta\{G(\beta) S(\beta)\}, \end{aligned}$$

so that $E_\beta\{G(\beta)S(\beta)\} = -E_\beta\{\partial G(\beta)/\partial \beta\}$, for all β . In particular, when $\beta = \beta_0$,

$$E\{G(\beta_0)S(\beta_0)\} = -E \left\{ \frac{\partial G(\beta)}{\partial \beta} \Big|_{\beta=\beta_0} \right\} = -C(\beta_0).$$

Thus, it follows that

$$\begin{aligned} \text{cov}\{G(\beta_0), S(\beta_0)\} &= E\{G(\beta_0)S(\beta_0)'\} - E\{G(\beta_0)\}E\{S(\beta_0)\}' \\ &= E\{G(\beta_0)S(\beta_0)'\} = -C(\beta_0)', \end{aligned} \quad (\text{A.5})$$

since $E\{S(\beta_0)\} = \mathbf{0}$. From (A.3), the approximate variance of G is given by

$$\begin{aligned}
\text{var}\{G(\hat{\beta})\} &\approx \text{var}\{G(\beta_0)\} + 2\text{cov}\{G(\beta_0), C(\beta_0)' I^{-1}(\beta_0) S(\beta_0)\} \\
&\quad + \text{var}\{C(\beta_0)' I^{-1}(\beta_0) S(\beta_0)\} \\
&= \text{var}\{G(\beta_0)\} - 2C(\beta_0)' I^{-1}(\beta_0) C(\beta_0) + C(\beta_0)' I^{-1}(\beta_0) C(\beta_0) \\
&= \text{var}\{G(\beta_0)\} - C(\beta_0)' I^{-1}(\beta_0) C(\beta_0),
\end{aligned}$$

where the first equality follows from (A.5) and the fact that $\text{cov}\{S(\beta_0)\} = \mathcal{Q}(\beta_0)$. To derive an explicit (approximate) expression for $\text{var}\{G(\hat{\beta})\}$, we need to calculate $\text{var}\{G(\beta_0)\}$, $C(\beta_0)$, and the information matrix $\mathcal{Q}(\beta_0)$. Define the $n \times 1$ vector

$$\mathbf{a}(\beta) = \left(\frac{1 - 2p_1(\beta)}{p_1(\beta)\{1 - p_1(\beta)\}}, \frac{1 - 2p_2(\beta)}{p_2(\beta)\{1 - p_2(\beta)\}}, \dots, \frac{1 - 2p_n(\beta)}{p_n(\beta)\{1 - p_n(\beta)\}} \right)',$$

the $(p + 1) \times n$ matrix

$$\mathbf{Q}(\beta) = \left(\frac{\partial p_1(\beta)}{\partial \beta} \frac{\partial p_2(\beta)}{\partial \beta} \dots \frac{\partial p_n(\beta)}{\partial \beta} \right),$$

and the $n \times n$ matrix

$$\mathbf{D}(\beta) = \text{diag} \left[\frac{1}{p_1(\beta)\{1 - p_1(\beta)\}}, \frac{1}{p_2(\beta)\{1 - p_2(\beta)\}}, \dots, \frac{1}{p_n(\beta)\{1 - p_n(\beta)\}} \right].$$

Using this notation, it is straightforward to show that $\text{var}\{G(\beta_0)\} = \mathbf{1}' \mathbf{D}(\beta_0) \mathbf{1} - 4n$, $C(\beta_0) = -\mathbf{Q}(\beta_0) \mathbf{a}(\beta_0)$, and $\mathcal{Q}(\beta_0) = \mathbf{Q}(\beta_0) \mathbf{D}(\beta_0) \mathbf{Q}(\beta_0)'$, where $\mathbf{1}$ is an $n \times 1$ vector of ones.

APPENDIX B

The Fisher information matrix in Section 3.2 is given by

$$\begin{aligned}
I(\theta) &= E\{S(\theta) S(\theta)'\} \\
&= E \left(\left[\sum_{i=1}^n \frac{Z_i - p_i(\theta)}{p_i(\theta)\{1 - p_i(\theta)\}} \left\{ \frac{\partial p_i(\theta)}{\partial \theta} \right\} \right] \left[\sum_{i=1}^n \frac{Z_i - p_i(\theta)}{p_i(\theta)\{1 - p_i(\theta)\}} \left\{ \frac{\partial p_i(\theta)}{\partial \theta} \right\} \right]' \right) \\
&= \sum_{i=1}^n E \left[\frac{\{Z_i - p_i(\theta)\}^2}{p_i(\theta)^2 \{1 - p_i(\theta)\}^2} \left\{ \frac{\partial p_i(\theta)}{\partial \theta} \frac{\partial p_i(\theta)}{\partial \theta'} \right\} \right] \\
&= \sum_{i=1}^n \left[\frac{1}{p_i(\theta)\{1 - p_i(\theta)\}} \left\{ \frac{\partial p_i(\theta)}{\partial \theta} \frac{\partial p_i(\theta)}{\partial \theta'} \right\} \right],
\end{aligned}$$

where the penultimate equality holds because Z_i are independent and $E\{Z_i - p_i(\theta)\} = 0$. Substituting $\theta = \hat{\theta}$ into the last expression and simplifying, we obtain

$$\mathcal{I}(\widehat{\theta}) = \gamma_{12}^2 \sum_{i=1}^n \left[\left\{ \prod_{j=1}^{c_i} (1 - \widehat{p}_{ij}) \right\}^2 \frac{\mathbf{w}_i \mathbf{w}_i'}{\widehat{p}_i (1 - \widehat{p}_i)} \right],$$

$$\text{where } \mathbf{w}_i = \sum_{j=1}^{c_i} \widehat{p}_{ij}(\mathbf{x}_{ij}', \widehat{\eta}_{ij}^2 I(\widehat{\eta}_{ij} < 0)/2)',$$

Table I

Simulation study for the omission of the quadratic term in model (8) with $a = 0.0014$, 0.01 , 0.03 , 0.05 , 0.10 , and $\gamma_1 = \gamma_2 = 1$. The null model corresponds to $a = 0.0014$ and the effect of misspecification becomes more pronounced as a increases. The nominal size is 0.05 . Estimated sizes and powers are based on 500 simulated group testing data sets, each consisting of $N = 1000$ individuals, for different values of c (pool size) and n (number of pools) with $N = nc$. The margin of error for the estimated size, assuming a 99 percent confidence level, is 0.03 . RP (HP) denotes random (homogeneous) pooling.

c	n		RP					HP				
			$a = 0.0014$	0.01	0.03	0.05	0.10	$a = 0.0014$	0.01	0.03	0.05	0.10
2	500	G	0.05	0.31	0.62	0.79	0.94	0.06	0.52	0.88	0.98	1.00
		T_S	0.04	0.25	0.56	0.75	0.93	0.04	0.47	0.88	0.97	1.00
		T_{HL}	0.05	0.13	0.25	0.34	0.46	0.05	0.34	0.66	0.84	0.98
		IOS_A	0.05	0.28	0.64	0.78	0.95	0.06	0.49	0.90	0.97	1.00
5	200	G	0.04	0.14	0.27	0.36	0.56	0.06	0.37	0.85	0.97	1.00
		T_S	0.06	0.10	0.20	0.28	0.46	0.06	0.34	0.84	0.96	1.00
		T_{HL}	0.06	0.06	0.07	0.07	0.08	0.06	0.20	0.50	0.74	0.90
		IOS_A	0.03	0.14	0.25	0.35	0.53	0.04	0.32	0.74	0.94	1.00
10	100	G	0.04	0.07	0.10	0.11	0.14	0.05	0.23	0.56	0.75	0.86
		T_S	0.05	0.06	0.07	0.09	0.13	0.05	0.27	0.63	0.85	0.98
		T_{HL}	0.05	0.04	0.04	0.05	0.07	0.03	0.16	0.34	0.48	0.78
		IOS_A	0.05	0.11	0.13	0.15	0.23	0.06	0.05	0.20	0.29	0.71

Simulation study for the omission of the interaction term in model (9) with $\beta_3 = 0, 0.3, 0.5, 0.8, 1.0$, and $\gamma_1 = \gamma_2 = 1$. The null model corresponds to $\beta_3 = 0$ and the effect of misspecification becomes more pronounced as β_3 increases. The nominal size is 0.05. Estimated sizes and powers are based on 500 simulated group testing data sets, each consisting of $N = 1000$ individuals, for different values of c (pool size) and n (number of pools) with $N = nc$. The margin of error for the estimated size, assuming a 99 percent confidence level, is 0.03. RP (HP) denotes random (homogeneous) pooling.

Table II

c	n		RP					HP				
			$\beta_3 = 0$	0.3	0.5	0.8	1.0	$\beta_3 = 0$	0.3	0.5	0.8	1.0
2	500	G	0.05	0.22	0.45	0.78	0.87	0.05	0.35	0.74	0.99	1.00
		T_S	0.05	0.16	0.33	0.63	0.81	0.04	0.28	0.67	0.97	1.00
		T_{HL}	0.04	0.10	0.14	0.35	0.54	0.04	0.18	0.49	0.88	0.97
		IOS_A	0.04	0.03	0.05	0.10	0.26	0.04	0.06	0.11	0.27	0.53
5	200	G	0.08	0.15	0.24	0.36	0.46	0.07	0.35	0.68	0.97	0.99
		T_S	0.05	0.09	0.11	0.17	0.22	0.04	0.27	0.62	0.96	0.99
		T_{HL}	0.05	0.04	0.06	0.08	0.07	0.05	0.17	0.38	0.80	0.95
		IOS_A	0.06	0.05	0.06	0.05	0.08	0.05	0.03	0.07	0.11	0.31
10	100	G	0.04	0.08	0.09	0.13	0.14	0.05	0.17	0.46	0.88	0.93
		T_S	0.07	0.06	0.08	0.07	0.08	0.05	0.20	0.54	0.92	0.99
		T_{HL}	0.05	0.05	0.04	0.05	0.04	0.05	0.12	0.29	0.62	0.87
		IOS_A	0.05	0.05	0.03	0.04	0.03	0.06	0.03	0.02	0.04	0.07

Table III

Simulation study for the misspecification of the link function in model (10) with $\alpha_0 = 0, 0.5, -0.5, -1.0$, and $\gamma_1 = \gamma_2 = 1$. The null model corresponds to $\gamma_0 = 0$. The nominal size is 0.05. Estimated sizes and powers are based on 500 simulated group testing data sets, each consisting of $N = 1000$ individuals, for different values of c (pool size) and n (number of pools) with $N = nc$. The margin of error for the estimated size, assuming a 99 percent confidence level, is 0.03. Cases with “-” represent situations where complete covariate separation occurs too often. RP (HP) denotes random (homogeneous) pooling.

c	n		RP				HP			
			$\alpha_0 = 0$	0.5	-0.5	-1	$\alpha_0 = 0$	0.5	-0.5	-1
2	500	G	0.06	0.04	0.43	0.48	0.06	0.24	0.56	0.60
		T_S	0.04	0.07	0.52	0.45	0.04	0.07	0.64	0.58
		T_{HL}	0.05	0.01	0.29	0.14	0.05	0.00	0.45	0.32
		IOS_A	0.06	0.00	0.50	0.44	0.05	0.02	0.64	0.57
5	200	G	0.06	0.01	0.21	0.15	0.04	0.55	0.42	0.27
		T_S	0.05	0.04	0.21	0.17	0.03	0.05	0.48	0.39
		T_{HL}	0.04	0.04	0.07	0.05	0.04	0.02	0.29	0.17
		IOS_A	0.06	0.00	0.26	0.21	0.05	0.09	0.34	0.21
10	100	G	0.04	0.01	0.08	0.06	0.08	-	0.22	0.01
		T_S	0.05	0.04	0.10	0.07	0.04	-	0.28	0.13
		T_{HL}	0.05	0.04	0.10	0.07	0.04	-	0.13	0.08
		IOS_A	0.02	0.02	0.09	0.11	0.06	-	0.03	0.02

Kenya HIV data. Parameter estimates and standard errors from the model in (11) for chronological and ordered pooling strategies with $\gamma_1 = 1$ and $\gamma_2 = 0.9997$. Probability values for GOF are also provided. Results are based on 100 pools of size 7 and 1 pool of size 6.

Table IV

p-value						
Pooling strategy	β_0	β_1	G	T_S	T_{HL}	IOS_A
Chronological (RP)	-2.02(1.43)	-0.02(0.06)	0.84	0.30	0.99	0.84
Ordered (HP)	-2.01(0.46)	-0.02(0.02)	0.80	0.24	0.28	0.86

Table V

Nebraska IPP data. GOF testing results for the model in (12) with $\gamma_1 = 0.98$ and $\gamma_2 = 0.95$. Results are based on 613 pools of size 10 and 1 pool of size 8.

Infection	p-value			
	G	T_S	T_{HL}	IOS_A
Chlamydia	0.05	0.08	0.89	0.06
Gonorrhea	0.19	0.71	0.25	0.15