Biometrika (2014), **101**, 3, *pp.* 587–598
Printed in Great Britain
doi: 10.1093/biomet/asu007
Advance Access publication 13 May 2014

Semiparametric group testing regression models

By D. WANG, C. S. McMAHAN, C. M. GALLAGHER

Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29631, U.S.A.

dwang@g.clemson.edu mcmaha2@clemson.edu cgallag@clemson.edu

and K. B. KULASEKERA

Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, Kentucky 40202, U.S.A.

kb.kulasekera@louisville.edu

Summary

Group testing, through the use of pooling, has proven to be an efficient method of reducing the time and cost associated with screening for a binary characteristic of interest, such as infection status. A topic of key interest in the statistical literature involves the development of regression models that relate individual-level covariates to testing responses observed from pooled specimens. In this article, we propose a general semiparametric framework that allows for the inclusion of multi-dimensional covariates, decoding information, and imperfect testing. The asymptotic properties of our estimators are presented and guidance on finite sample implementation is provided. We illustrate the performance of our methods through simulation and by applying them to chlamydia and gonorrhea data collected by the Nebraska Public Health Laboratory as a part of the Infertility Prevention Project.

Some key words: Group testing; Latent data; Pooled data; Semiparametric regression; Sensitivity; Specificity.

1. Introduction

Group testing, also known as pooled testing, was first proposed by Dorfman (1943) as a means to reduce the cost associated with screening World War II inductees for syphilis. In order to reduce testing expenditure, Dorfman suggested that pooled specimens, formed from combining blood samples collected from individuals, be tested for the presence of syphilis. If the initial pool, also referred to as a master pool, tested negative, then all contributing men could be declared negative at the cost of only one test. Alternatively, positive master pools would be resolved by retesting each of the contributing specimens one by one. Since this seminal work, many variants of Dorfman's decoding strategy have been proposed in an effort to further reduce screening costs or increase classification accuracy; for a review see Kim et al. (2007).

In addition to being used for case identification, pooling techniques have also been implemented for the purposes of estimation, predominantly in the context of estimating population level characteristics; see Bilder & Tebbs (2005) for a review. More recently, authors have developed binary regression models that relate pool response data to individual-level covariate information through a specified link function; see Vansteelandt et al. (2000), Bilder & Tebbs (2009), Chen et al. (2009), and Huang & Tebbs (2009). To obviate the specification of the link function, Delaigle & Meister (2011), Delaigle & Hall (2012), and Wang et al. (2013) proposed

nonparametric binary regression techniques for group testing data that allow for the incorporation of a single continuous explanatory variable. Delaigle & Meister (2011) discussed extensions of their approach that allow for multiple covariates via a multivariate kernel function. However, due to the curse of dimensionality this approach may not be suitable for evaluating multiple explanatory variables. The aforementioned regression methods were designed to model data arising from master pool testing only; i.e., these methods cannot incorporate information gained from decoding positive pools. To our knowledge, the only binary regression models that allow for the incorporation of decoding information were proposed by Xie (2001) and Zhang et al. (2013), and were developed under parametric assumptions.

Since its advent, group testing has been successfully implemented for screening for a variety of infectious diseases (Lewis et al., 2012; Van et al., 2012), and has found applications in areas such as genetics (Gastwirth, 2000), drug discovery (Remlinger et al., 2006), medical entomology (Venette et al., 2002), veterinary science (Muñoz-Zanzi et al., 2000), and plant pathology (Venette et al., 2002). The group testing strategy implemented varies according to the goals of the study and often does not conclude with master pool testing. Consequently, in this paper we propose a general regression methodology for modelling test responses obtained from all group testing algorithms that allows for the incorporation of multiple covariates and accounts for imperfect testing. Unlike the aforementioned parametric methods, our semiparametric model enjoys the modelling flexibility of nonparametric procedures, but is not subject to the curse of dimensionality when multiple predictors are available. We develop hypothesis-testing methods for evaluating the significance of potential predictors based on the asymptotic properties of our proposed estimators. Through simulation, we illustrate that our methodology can more reliably evaluate potential predictors when compared to analogous parametric methods.

Our methodology falls broadly into the class of single-index models, which have attracted much attention in the statistical literature over the past few decades; see Ichimura (1993), Härdle et al. (1993), Klein & Spady (1993), Xia et al. (2002), Xia (2006), Zhu & Xue (2006), Cui et al. (2011) and the references therein. Though similar, there exists a fundamental difference between our method and those previously proposed in the literature. Specifically, all existing single-index models require that a response be available for each individual, while in contrast our method requires only the availability of the responses obtained from testing pools of individuals. Therefore, the complex data structure resulting from group testing algorithms cannot be handled by any of the existing single-index techniques.

2. MODELS AND METHODOLOGY

2·1. *Modelling assumptions and general estimation procedure*

In what follows, we propose a general modelling framework for data arising from any group testing algorithm. Our proposed methodology can be greatly simplified under two of the most common such algorithms, master pool testing and Dorfman decoding, as is illustrated in the subsequent sections. Consider implementing a group testing algorithm to screen N individuals for a binary characteristic of interest, such as infection status. In general, this process begins by randomly assigning each of the individuals to exactly one of J initial groups of size c_j. Let $\mathcal{G}_j = \{1, \ldots, c_j\}$ be a collection of indices identifying the c_j individuals assigned to the jth group. Within the jth group, screening is performed according to the protocol outlined by the specified group testing algorithm, resulting in K_j testing responses Y_{jl}, for $l = 1, \ldots, K_j$. We let $Y_{jl} = 1$ indicate that the lth pool tested positive, and $Y_{jl} = 0$ otherwise. We identify the individuals in the jth group whose specimens were pooled and tested by the lth assay by the set $\mathcal{P}_{jl} \subseteq \mathcal{G}_j$, and we

define $Z_{jl} = (Y_{jl}, \mathcal{P}_{jl})$. For notational convenience, we collect all of the observed testing data associated with the jth group into the set $Z_j = \{Z_{j1}, \ldots, Z_{jK_j}\}$, and we assume throughout that $Z_j \perp\!\!\!\perp Z_{j'}$ for all $j \neq j'$, where $\perp\!\!\!\perp$ denotes statistical independence.

Let T_{ij} denote the true status of the ith individual in the jth group, where $T_{ij} = 1$ indicates that the individual is positive, and $T_{ij} = 0$ otherwise. For modelling purposes, we assume that $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^{\mathrm{T}}$, a p-dimensional vector of covariates, is available for each individual and that the random vectors (T_{ij}, X_{ij}) are independent and identically distributed. In order to relate the individuals' true statuses to their predictor variables, we proceed under the single-index generalization; i.e., we assume that $\mathrm{pr}(T_{ij} = 1 \mid X_{ij} = x) = p(x^{\mathrm{T}}\beta)$, where $p(\cdot)$ is an unknown smooth probability curve and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional vector of regression parameters. To ensure identifiability, as with all single-index models, we assume that the support of the covariate vectors, $\mathbb{X}$, is a bounded convex set with at least one interior point and the parameter space of β is $\mathcal{B} = \{\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} : \|\beta\| = 1, \beta_1 > 0\}$, where $\|\beta\|$ denotes the Euclidean norm of β (Lin & Kulasekera, 2007). If one observed T_{ij}, for $i = 1, \ldots, c_j$ and $j = 1, \ldots, J$, then standard single-index estimation procedures could be employed to estimate $p(\cdot)$ and β, but when the assay being used is imperfect and the testing responses are based on pooled assessments the individuals' true statuses are latent and these techniques are inapplicable.

To account for imperfect testing, we let S_e and S_p denote the sensitivity and specificity of the assay being employed; i.e., S_e is the probability that a specimen will test positive given it is truly positive and S_p is the probability that a specimen will test negative given it is truly negative. We assume that S_e and S_p are known, constant, and independent of the pool size. Further, we assume that given the true status of the pools being tested, $Y_{jl} \perp\!\!\!\perp Y_{jl'}$, for $l \neq l'$. These assumptions are common in the group testing literature; see Xie (2001), Kim et al. (2007), and Zhang et al. (2013).

Using the testing error rates and these assumptions we now relate the observed testing outcomes to the true underlying statuses of the specimens being tested. To accomplish this, we let $\mathcal{Z}(c)$ denote the set of all possible outcomes resulting from screening a group of size c according to a specific group testing algorithm. Likewise, we define the set of all possible true statuses for the individuals assigned to a group of size c to be $\mathcal{T}(c)$. The conditional probability of observing any $Z = \{(Y_1, \mathcal{P}_1), \ldots, (Y_K, \mathcal{P}_K)\} \in \mathcal{Z}(c)$ given any $T = (T_1, \ldots, T_c) \in \mathcal{T}(c)$ can be calculated as

$$M(Z, T, c) = \mathrm{pr}(\mathcal{P}) \prod_{l=1}^{K} \left\{ S_e^{Y_l \tilde{Y}_l} (1 - S_e)^{(1-Y_l)\tilde{Y}_l} (1 - S_p)^{Y_l(1-\tilde{Y}_l)} S_p^{(1-Y_l)(1-\tilde{Y}_l)} \right\},$$

where $\tilde{Y}_l = \max_{i \in \mathcal{P}_l} T_i$ and $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$. The probability $\mathrm{pr}(\mathcal{P})$ accounts for the randomness, if any, in the pooling protocol of the group testing algorithm. In the Supplementary Material we provide a derivation of $M(Z, T, c)$ and illustrate how $\mathrm{pr}(\mathcal{P})$ should be evaluated.

In what follows we relate the observed testing outcomes arising from a group testing algorithm to the individual-level covariate information. Through an application of the law of total probability it is easy to show that the conditional probability of observing Z_j given β, $p(\cdot)$, and $\mathcal{X}_j$ can be expressed as

$$\mathcal{R}\{Z_j; \mathcal{X}_j, \beta, p(\cdot)\} = \sum_{T \in \mathcal{T}(c_j)} M(Z_j, T, c_j) \prod_{i=1}^{c_j} p(X_{ij}^{\mathrm{T}}\beta)^{T_i} \{1 - p(X_{ij}^{\mathrm{T}}\beta)\}^{1-T_i}, \tag{1}$$

where $\mathcal{X}_j = (X_{1j}, \ldots, X_{c_j j})^{\mathrm{T}}$. To derive (1) we proceed under the assumption that the observed testing outcomes are independent of the measured covariates, given the individuals' true statuses.

Thus, the full conditional loglikelihood of $\{(Z_1, \mathcal{X}_1), \ldots, (Z_J, \mathcal{X}_J)\}$ can be expressed as

$$l\{\beta, p(\cdot)\} = \sum_{j=1}^{J} \log \mathcal{R}\{Z_j; \mathcal{X}_j, \beta, p(\cdot)\}.$$

If $p(\cdot)$ were known, an estimate of β could be obtained as the maximizer of $l\{\beta, p(\cdot)\}$. Thus, the primary challenge of fitting our model is to account for the dependence between the infinite-dimensional parameter $p(\cdot)$ and the finite-dimensional parameter β. To explicitly acknowledge this dependence, we write $p(\cdot)$ as $p_\beta(\cdot)$, and again point out that an estimate of β could be obtained as the maximizer of $l\{\beta, p_\beta(\cdot)\}$, if $p_\beta(\cdot)$ were known. In order to estimate the regression parameters, we propose to replace the unknown function $p_\beta(\cdot)$ by a consistent estimator, $\hat{p}_\beta(\cdot)$, so that our estimator of β can be obtained as $\hat{\beta} = \text{argmax}_{\beta \in \mathcal{B}}\, l\{\beta, \hat{p}_\beta(\cdot)\}$.

As previously stated, traditional single-index techniques are not applicable in this context, because the individuals' statuses are latent. To circumvent this, we propose to make use of the individuals' diagnosed statuses. To this end, let D_{ij} denote the diagnosed status of the ith individual in the jth group, such that $D_{ij} = 1$ indicates a positive diagnosis, and $D_{ij} = 0$ otherwise. Typically, an individual's diagnosed status is determined based on the observed testing outcomes and the specified testing protocol; i.e., $D_{ij} = \Lambda(i, Z_j)$, where Λ is a decision function unique to the group testing algorithm being implemented. Define $\mathcal{F}_{ij}(t, \mu) = \text{pr}(D_{ij} = 1 \mid T_{ij} = t)$, which can be calculated as

$$\mathcal{F}_{ij}(t, \mu) = \sum_{Z \in \mathcal{Z}_i(c_j)} \sum_{T \in \mathcal{T}(c_j)} I(T_i = t) M(Z, T, c_j) \prod_{k \neq i} \{\mu^{1-T_k}(1-\mu)^{T_k}\},$$

where $\mu = \text{pr}(T_{ij} = 0)$ and $\mathcal{Z}_i(c) = \{z \in \mathcal{Z}(c) : \Lambda(i; z) = 1\}$; i.e., $\mathcal{Z}_i(c)$ is the set of all possible testing outcomes that would result in the ith individual in a group of size c being diagnosed positive. The quantities $\mathcal{F}_{ij}(1, \mu)$ and $1 - \mathcal{F}_{ij}(0, \mu)$ are commonly referred to as the pooling sensitivity and specificity, respectively, and under specific group testing algorithms these measures of testing accuracy have nice analytic forms; see Kim et al. (2007).

In order to develop an estimator of $p_\beta(\cdot)$, we consider the conditional probability that an individual will be diagnosed positive, given the linear predictor $X_{ij}^{\mathrm{T}}\beta$, which can be expressed as

$$E(D_{ij} \mid X_{ij}^{\mathrm{T}}\beta = u) = a_{ij}(\mu) + b_{ij}(\mu) p_\beta(u), \tag{2}$$

where $a_{ij}(\mu) = \mathcal{F}_{ij}(0, \mu)$ and $b_{ij}(\mu) = \mathcal{F}_{ij}(1, \mu) - \mathcal{F}_{ij}(0, \mu)$. The unknowns in (2) are μ and $p_\beta(\cdot)$. Since μ is the unconditional probability that an individual is truly negative, one could obtain an estimator, $\hat{\mu}$, of this parameter by maximizing the full loglikelihood

$$l_p(\mu) = \sum_{j=1}^{J} \log \left(\sum_{T \in \mathcal{T}(c_j)} \left[M(Z_j, T, c_j) \prod_{i=1}^{c_j} \{\mu^{1-T_i}(1-\mu)^{T_i}\} \right] \right), \tag{3}$$

with respect to μ; i.e., $\hat{\mu} = \text{argmax}_\mu l_p(\mu)$. Then, based on equation (2), we can obtain a local linear kernel estimator of $p_\beta(\cdot)$ at a given point u by minimizing

$$\sum_{j=1}^{J} \sum_{i=1}^{c_j} \left[D_{ij} - a_{ij}(\hat{\mu}) - b_{ij}(\hat{\mu}) \{p_\beta(u) + p'_\beta(u)(X_{ij}^{\mathrm{T}}\beta - u)\} \right]^2 K_h(X_{ij}^{\mathrm{T}}\beta - u), \tag{4}$$

with respect to $\{p_\beta(u), p'_\beta(u)\}^{\mathrm{T}}$, where $p'_\beta(\cdot)$ denotes the first derivative of $p_\beta(\cdot)$, h is a user defined bandwidth, $K(\cdot)$ is a symmetric kernel density function, and $K_h(\cdot) = h^{-1}K(\cdot/h)$. We define $\{\hat{p}_\beta(u), \hat{p}'_\beta(u)\}^{\mathrm{T}}$, the minimizer of (4), to be our estimator of $\{p_\beta(u), p'_\beta(u)\}^{\mathrm{T}}$. Consequently, our final estimators can be expressed as

$$\hat{\beta} = \mathrm{argmax}_{\beta\in\mathcal{B}}\, l\left\{\beta, \hat{p}_\beta(\cdot)\right\}, \quad \hat{p}(u) = \hat{p}_{\hat{\beta}}(u). \tag{5}$$

For computational reasons, $\hat{p}_\beta(u)$ can be expressed in closed form, but this expression is omitted for brevity. Further, (4) is not the standard form of the local sum of squares, because the diagnosed statuses are correlated and $\hat{\mu}$ is a random term that depends on the observed testing data. Despite these differences, in § 3 we show that our approach efficiently estimates β and $p(\cdot)$. In the following two sections, we outline the formulas necessary to implement our regression methodology under master pool testing and Dorfman decoding. A more detailed illustration is provided in the Supplementary Material.

2·2. *Estimation under master pool testing*

The testing protocol under master pool testing specifies that specimens collected from individuals belonging to a common group be combined to form a single master pool that is subsequently assayed; i.e., the testing data available for modelling are $Z_j = \{(Y_{j1}, \mathcal{P}_{j1})\}$, where $\mathcal{P}_{j1} = \mathcal{G}_j$. If $Y_{j1} = 0$, then all individuals in this group are diagnosed as negative, whereas $Y_{j1} = 1$ indicates that at least one individual is at risk. Thus, we define $D_{ij} = \Lambda(i, Z_j) = Y_{j1}$. Under master pool testing, the loglikelihood (3) reduces to

$$l_p(\mu) = \sum_{j=1}^{J} (1 - Y_{j1}) \log p_{j0} + Y_{j1} \log(1 - p_{j0}),$$

where $p_{j0} = 1 - S_e - \delta_{c_j}$ and $\delta_c = (1 - S_e - S_p)\mu^c$. Similarly, a series of simple arguments provide that $a_{ij}(\mu) = S_e + \delta_{c_j - 1}$ and $b_{ij}(\mu) = S_e - a_{ij}(\mu)$. Finally, for the jth group the observed testing data Z_j belong to the set $\{(0, \mathcal{P}_{j1}), (1, \mathcal{P}_{j1})\}$, and the conditional probability outlined in (1) associated with either of these outcomes is $\mathcal{R}\{(0, \mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\} = 1 - S_e - \delta_0 \prod_{i=1}^{c_j} \{1 - p(X_{ij}^{\mathrm{T}}\beta)\}$ or $\mathcal{R}\{(1, \mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\} = 1 - \mathcal{R}\{(0, \mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\}$. The estimators defined in (5) are then obtained as described in § 2·1.

2·3. *Estimation under Dorfman decoding*

Dorfman decoding proceeds in a similar fashion to master pool testing, with the key difference that positive pools are resolved by retesting all contributing individuals one by one. Consequently, Z_j can take two forms, the first being $Z_j = \{(Y_{j1}, \mathcal{P}_{j1})\}$, where $Y_{j1} = 0$ and $\mathcal{P}_{j1} = \mathcal{G}_j$, denoting that the master pool tested negative. The second occurs when the master pool test is positive; i.e., $Y_{j1} = 1$ and $\mathcal{P}_{j1} = \mathcal{G}_j$, in which case $Z_j = \{(Y_{j1}, \mathcal{P}_{j1}), \ldots, (Y_{jK_j}, \mathcal{P}_{jK_j})\}$ where $K_j = c_j + 1$ and $\mathcal{P}_{jl} = \{l - 1\}$, for $l = 2, \ldots, K_j$. The ith individual's diagnosed status is determined to be $D_{ij} = \Lambda(i, Z_j) = 1$ if and only if $Y_{j1} = 1$ and $Y_{j,i+1} = 1$, $D_{ij} = \Lambda(i, Z_j) = 0$ otherwise; i.e., a positive diagnosis requires both the master pool and individual-level test to be positive. Under Dorfman testing, the loglikelihood (3) reduces to

$$l_p(\mu) = \sum_{j=1}^{J} \left\{ I(Y_{j1} = 0) \log p_{j0} + \sum_{k=0}^{c_j} I\left(Y_{j1} = 1, \sum_{l=2}^{c_j+1} Y_{jl} = k\right) \log p_{j1k} \right\}.$$

where $p_{j1k} = \delta_{c_j}(1-S_p)^k S_p^{c_j-k} + S_e(S_e+\delta_1)^k(1-S_e-\delta_1)^{c_j-k}$, $p_{j0} = 1-S_e-\delta_{c_j}$, and $\delta_c = (1-S_e-S_p)\mu^c$. Similarly, simple arguments yield $a_{ij}(\mu) = (1-S_p)^2\mu^{c_j-1} + S_e(1-S_p)(1-\mu^{c_j-1})$ and $b_{ij}(\mu) = S_e^2 - a_{ij}(\mu)$.

The approach described in § 2·2 can be used to calculate the probability that the jth master pool will test negative; i.e., in this case we have that $\mathcal{R}\{(0, \mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\} = 1 - S_e - \delta_0 \prod_{i=1}^{c_j}\{1 - p(X_{ij}^{\mathrm{T}}\beta)\}$. To express the probability of the other testing outcomes, we define $\mathcal{I}_{j1} = \{i \in \mathcal{G}_j : D_{ij} = 1\}$ and $\mathcal{I}_{j0} = \{i \in \mathcal{G}_j : D_{ij} = 0\}$; i.e., the sets $\mathcal{I}_{j1}$ and $\mathcal{I}_{j0}$ identify the $k = |\mathcal{I}_{j1}|$ and $c_j - k = |\mathcal{I}_{j0}|$ individuals in the jth group that were diagnosed as positive and negative, respectively. Thus, for other testing outcomes $\mathcal{R}\{Z_j; \mathcal{X}_j, \beta, p(\cdot)\}$ is

$$\sum_{k_1=0}^{k}\sum_{k_0=0}^{c_j-k} S_e^{k_1+I(k_1+k_0>0)}(1-S_e)^{k_0} S_p^{c_j-k-k_0}(1-S_p)^{k-k_1+I(k_1+k_0=0)} \prod_{l=0}^{1} \mathrm{pr}(\mathcal{S}_{jl} = k_l), \qquad (6)$$

where $\mathcal{S}_{jl} = \sum_{i\in\mathcal{I}_{jl}} T_{ij}$. The probabilities in (6) are conditional on the unknown parameters and predictor variables, so $\mathcal{S}_{j1}$ and $\mathcal{S}_{j0}$ are the sum of independent and nonidentically distributed Bernoulli random variables; i.e., $\mathcal{S}_{j1}$ and $\mathcal{S}_{j0}$ each follow a Poisson binomial distribution. The estimators defined in (5) are then obtained as described in § 2·1.

3. ASYMPTOTIC PROPERTIES

We assume that $J \to \infty$ as $N \to \infty$ while group sizes remain finite. This is reasonable since in practice the group sizes are naturally bounded by implementation considerations. Further, this assumption is common in the group testing literature; see Delaigle & Meister (2011). We denote the range of c_j by $\{c^{(1)}, \ldots, c^{(M)}\}$. More explicitly, for all pooled observations there exists an m such that $c_j = c^{(m)}$. Further, for each m we let J_m denote the number of groups having size $c^{(m)}$, and assume that $J_m c^{(m)}/N \to \gamma_m$ as $N \to \infty$; i.e., γ_m represents the proportion of individuals assigned to groups of size $c^{(m)}$.

Theorem 1 provides the asymptotic properties of our proposed estimators $\hat{\beta}$ and $\hat{p}(\cdot)$. In order to succinctly present these results we let $\beta_0 = (\beta_{01}, \beta_0^{(1)\mathrm{T}})^{\mathrm{T}}$ and $p_0(\cdot)$ denote the true unknown parameters, where $\beta_0^{(1)} = (\beta_{02}, \ldots, \beta_{0p})^{\mathrm{T}}$. We define

$$\Omega_c = c^{-1}\sum_{z\in\mathcal{Z}(c)} E\left[\mathcal{R}^{-1}\{z; \mathcal{X}^{(c)}, \beta_0, p_0(\cdot)\}\sum_{i=1}^{c}\{P_i(z,1,c) - P_i(z,0,c)\}^2 p_0'^2(X_i^{\mathrm{T}}\beta_0)\Gamma(X_i)\right],$$

where $\mathcal{X}^{(c)} = (X_1, \ldots, X_c)^{\mathrm{T}}$, $\Gamma(X) = \{X - E(X \mid X^{\mathrm{T}}\beta_0\}\{X - E(X^{\mathrm{T}} \mid X^{\mathrm{T}}\beta_0)\}^{\mathrm{T}}$, and $P_i(z,t,c) = \mathrm{pr}\{Z = z \mid T_i = t, \mathcal{X}^{(c)}, \beta_0, p_0(\cdot)\}$. Finally, we define $\Omega = \sum_{m=1}^{M}\gamma_m\Omega_{c^{(m)}}$, which plays an integral role in the asymptotic variance covariance matrix of $\hat{\beta}$. Under a specific testing protocol, e.g., master pool testing or Dorfman decoding, the above expression for Ω can be more explicit. To illustrate this fact, in the Supplementary Material we provide distinct versions of Ω for the methodology described in § 2·2 and § 2·3. Using the above expressions we now give our main result.

THEOREM 1. *Under Conditions A1–A5 in the Appendix, we have that*

$$N^{1/2}(\hat{\beta} - \beta_0) \to N(0, \Sigma)$$

in distribution, where $\Sigma = \mathcal{J}_0(\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0)^{-1}\mathcal{J}_0^{\mathrm{T}}$, $\mathcal{J}_0$ *is the functional value of* $\partial B(\beta^{(1)})/\partial\beta^{(1)}$ *evaluated at* $\beta^{(1)} = \beta_0^{(1)}$, *and* $B(\beta^{(1)}) = ([1 - \|\beta^{(1)}\|^2]^{1/2}, \beta^{(1)\mathrm{T}})^{\mathrm{T}}$. *Further,*

$$\sup_{x\in\mathbb{X}} |\hat{p}(x^{\mathrm{T}}\hat{\beta}) - p_0(x^{\mathrm{T}}\beta_0)|^2 = O_p\left\{(\log N)/(Nh)\right\}.$$

The consistency rate for estimating $p_0(\cdot)$ is the same rate demonstrated for kernel smoothing estimators in a univariate nonparametric regression context; see Mack & Silverman (1982). The estimator $\hat{\mu}$ is a maximum likelihood estimator, its asymptotic normality follows from standard arguments and hence is omitted.

Theorem 1 suggests that large sample inference is possible once a good estimator $\hat{\Sigma}$ of Σ is obtained. To this end, the Supplementary Material gives an extension of a plug-in estimator of Σ that was originally proposed by Wang et al. (2010). Using $\hat{\beta}$ and $\hat{\Sigma}$ one can conduct Wald type inference; i.e., at the significance level α, a confidence interval for β_{0r} can be constructed as

$$\hat{\beta}_{0r} \pm \Phi^{-1}(1 - \alpha/2)\hat{\sigma}_r N^{-1/2} \quad (r = 1, \ldots, p),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution and $\hat{\sigma}_r^2$ is the rth diagonal element of $\hat{\Sigma}$. Further, for $r < p$ one may also perform hypothesis tests of the form

$$H_0 : \beta_{0q_1} = \cdots = \beta_{0q_r} = 0 \quad \text{versus} \quad H_1 : \text{not all } \beta_{0q_1}, \ldots, \beta_{0q_r} \text{ equal } 0,$$

using the test statistic $R_N = N(D\hat{\beta})^{\mathrm{T}}(D\hat{\Sigma}D^{\mathrm{T}})^{-1}D\hat{\beta}$, where D is a $r \times p$ matrix such that $D\beta_0 = (\beta_{0q_1}, \ldots, \beta_{0q_r})^{\mathrm{T}}$. Given the results in Theorem 1, we have that under the null hypothesis R_N converges in distribution to a chi-square random variable having r degrees of freedom. Consequently, at the significance level α one would reject the null hypothesis if $R_N > \chi_r^2(1 - \alpha)$, where $\chi_r^2(a)$ is the ath quantile of a chi-square distribution having r degrees of freedom.

4. Numerical analysis

A simulation study was conducted to assess the finite sample performance of our methodology. This study considered the following three underlying true regression models:

Model 1: $p_0(u) = 1/\{1 + \exp(4 - 2u)\}$,
Model 2: $p_0(u) = \exp(-5u^2 - 1{\cdot}5)$,
Model 3: $p_0(u) = [\sin\{\pi(u - 0{\cdot}3)\} + 1{\cdot}3]/[10 + 20(u - 0{\cdot}3)^2\{\mathrm{sign}(u - 0{\cdot}3) + 1\}]$,

where $u = X^{\mathrm{T}}\beta_0$. Model 1 provides a situation under which a logistic link is appropriate, and Models 2 and 3 emulate the gonorrhea and chlamydia data studied in § 5. For each of the above models we considered a vector of predictors of the form $X = (X_1, X_2, X_3)^{\mathrm{T}}$, where X_1 follows a standard normal distribution, while X_2 and X_3 each follow a Bernoulli distribution with success probabilities 0·4 and 0·3, respectively. The regression parameters were specified to be $\beta_0 = (\beta_{01}, \beta_{02}, \beta_{03})^{\mathrm{T}} = \{1/3, (8/9 - \delta^2)^{1/2}, \delta\}^{\mathrm{T}}$, where $\delta = \{0, 0{\cdot}1, 0{\cdot}2, 0{\cdot}3, 0{\cdot}4\}$.

We set $N = 10\,000$ and considered a common group size $c_j = c$ for all $j = 1, \ldots, J$, where $J = N/c$ and $c \in \{1, 2, 5, 10\}$. The setting $c = 1$ corresponds to individual-level testing. In order to generate group testing data, we first generated individual-level data; i.e., for each of the N individuals we generated the pair (T_{ij}, X_{ij}). Specifically, the predictor vector X_{ij} was simulated according to the distributions described above and T_{ij} was subsequently determined according to a Bernoulli(p_{ij}) distribution, where $p_{ij} = p_0(X_{ij}^{\mathrm{T}}\beta_0)$. To create group testing data, we then simulated the screening of the N individuals according to both master pool testing and

Dorfman decoding, chosen due to their popularity. To allow for testing errors, we generated testing responses using $S_e = 0{\cdot}93$ and $S_p = 0{\cdot}99$. Under both master pool testing and Dorfman decoding, this data generating process was repeated 500 times for each model and configuration of (c, δ).

For each of the group testing datasets we estimated the regression parameter β_0 and the link function $p_0(\cdot)$ using the methodology outlined in § 2. To implement our approach we specified $K(\cdot)$ to be the Gaussian kernel, and selected the bandwidth in a similar fashion to the method proposed in Härdle et al. (1993). Specifically, the bandwidth $\tilde{h}$ was chosen such that $(\tilde{\beta}, \tilde{h})$ is the maximizer of $\text{CV}(\beta, h) = \sum_{j=1}^{J} \log \mathcal{R}\{Z_j; \mathcal{X}_j, \beta, \hat{p}_\beta^{(-j)}(\cdot)\}$, where $\hat{p}_\beta^{(-j)}(u)$ denotes the leave-one-out estimator of $p_\beta(u)$ obtained from minimizing (4) when the information pertaining to the jth pool is omitted. For comparative purposes, we also implemented the parametric methods proposed in Vansteelandt et al. (2000) and Zhang et al. (2013) for master pool testing and Dorfman decoding, respectively, under the assumption the link function is logistic.

Table 1 provides summary statistics of the 500 estimates of β_0 obtained by our methodology, across all considered models and settings of c, under Dorfman decoding, when $\delta = 0{\cdot}1$. Our approach exhibits little, if any, evidence of bias and the average standard errors are in agreement with the sample standard deviation of the parameter estimates. The empirical coverage probabilities for 95% confidence intervals are predominantly at their nominal level. Further, the parameter estimates obtained from analysing group testing data can be as, if not more, efficient than the estimates based on individual-level data; i.e., in most cases the estimators have smaller variances when $c > 1$. This suggests that more precise inference can be obtained from analysing group testing decoding data, when compared to individual-level testing information, and at a fraction of the cost of data collection; similar findings were reported in Zhang et al. (2013).

Table 1 also provides the average mean squared error of prediction, where we define $\text{MSE}\{\hat{\beta}, \hat{p}(\cdot)\} = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{c_j} \{\hat{p}(X_{ij}^{\mathrm{T}}\hat{\beta}) - p_0(X_{ij}^{\mathrm{T}}\beta_0)\}^2$ to be the mean squared error of prediction for a given dataset. This measure suggests that our methodology can more accurately estimate the link function, using decoding data, than the analogous method that makes use of individual-level testing information. Table 1 provides the ratio of the average mean squared error of prediction for the parametric and our semiparametric model. We see that when the true underlying model is logistic the average mean squared error of prediction of our approach is roughly three times larger than that of the parametric model, which assumes a logistic link. In contrast, when the true model is not logistic the average mean squared error of prediction associated with the parametric model can be up to thirty times greater than that of our methodology.

We conducted a power analysis of the hypothesis test for β_{03}, using the estimates resulting from our regression procedures and the methodology outlined in § 3 to perform the test of $H_0 : \beta_{03} = 0$ versus $H_1 : \beta_{03} \neq 0$, at the $\alpha = 0{\cdot}05$ significance level. The same analysis was also performed for each dataset using the aforementioned parametric models, again assuming a logistic link. The hypothesis-testing results were used to construct power curves for our semiparametric approach and the competing parametric model, across all considered configurations. The power curves corresponding to data arising from Dorfman decoding when $c = 5$ are presented in Fig. 1. Under both the semiparametric and parametric models the hypothesis-testing procedure suggested in § 3 maintains its correct size across all considered settings. The estimated power curves under Model 1 are very similar, with the parametric model having slightly more power. This suggests that our methodology performs almost as well as the parametric model, which assumes the correct link function. If the link function is misspecified under the parametric model these methods lose the power to detect significant predictor variables, a feature not shared by our approach.

The results presented in Table 1 and Fig. 1 are based on analysing data arising from Dorfman decoding, and the parameter estimates summarized in Table 1 correspond to the case in which

Table 1. *Summary of simulation results for data arising from Dorfman decoding*

	Parameter	Measure	$c=1$	$c=2$	$c=5$	$c=10$
Model 1	β_{01}	BIAS (SD)	8·7 (3·5)	9·0 (3·1)	6·5 (3·3)	7·1 (3·1)
		COV (SE)	93·6 (3·5)	94·4 (3·2)	95·3 (3·2)	95·8 (3·3)
	β_{02}	BIAS (SD)	−5·1 (1·4)	−4·7 (1·3)	−4·4 (1·3)	−4·4 (1·3)
		COV (SE)	96·0 (1·4)	96·2 (1·3)	96·2 (1·3)	96·8 (1·4)
	β_{03}	BIAS (SD)	−1·9 (5·2)	−4·2 (4·9)	−1·6 (5·4)	−4·1 (5·6)
		COV (SE)	94·4 (5·3)	94·6 (5·0)	93·2 (5·1)	92·6 (5·3)
	$p_0\ (x\beta_0)$	EMSE (RE)	1·31 (0·37)	1·25 (0·35)	1·28 (0·38)	1·27 (0·39)
	Percentage reduction in testing			37·3%	52·5%	43·6%
Model 2	β_{01}	BIAS (SD)	1·5 (1·4)	0·7 (1·4)	0·5 (1·4)	1·9 (1·4)
		COV (SE)	93·0 (1·4)	95·3 (1·4)	93·8 (1·4)	95·1 (1·4)
	β_{02}	BIAS (SD)	−1·2 (0·6)	−1·0 (0·6)	−0·6 (0·6)	−1·1 (0·6)
		COV (SE)	93·4 (0·6)	94·5 (0·6)	93·6 (0·6)	96·2 (0·6)
	β_{03}	BIAS (SD)	−0·7 (3·4)	1·2 (3·2)	−2·8 (3·2)	−2·6 (3·1)
		COV (SE)	93·0 (3·0)	92·3 (2·9)	92·6 (2·9)	92·9 (3·0)
	$p_0\ (x\beta_0)$	EMSE (RE)	1·25 (25·33)	1·09 (29·83)	1·18 (27·43)	1·18 (27·24)
	Percentage reduction in testing			31·9%	41·9%	29·7%
Model 3	β_{01}	BIAS (SD)	7·6 (2·5)	8·9 (2·4)	8·5 (2·4)	7·5 (2·4)
		COV (SE)	92·4 (2·5)	92·8 (2·4)	92·3 (2·5)	93·0 (2·5)
	β_{02}	BIAS (SD)	−3·7 (1·0)	−4·4 (1·0)	−4·3 (1·0)	−4·0 (1·0)
		COV (SE)	93·8 (1·0)	92·8 (1·0)	94·3 (1·0)	93·0 (1·0)
	β_{03}	BIAS (SD)	−1·7 (3·7)	−0·2 (3·9)	1·5 (3·6)	1·3 (4·0)
		COV (SE)	92·4 (3·6)	92·4 (3·5)	94·0 (3·6)	92·7 (3·6)
	$p_0\ (x\beta_0)$	EMSE (RE)	1·61 (13·80)	1·46 (15·18)	1·46 (15·19)	1·57 (14·19)
	Percentage reduction in testing			34·8%	47·4%	36·7%

BIAS and SD, empirical bias ($\times 10^3$) and standard deviation ($\times 100$) of the 500 estimates; SE, average standard error ($\times 100$); COV, empirical coverage probability ($\times 100$) for nominal 95% confidence interval; EMSE, average mean squared error of prediction ($\times 10^4$); RE, ratio of EMSE of the parametric model to the EMSE of our semiparametric model.

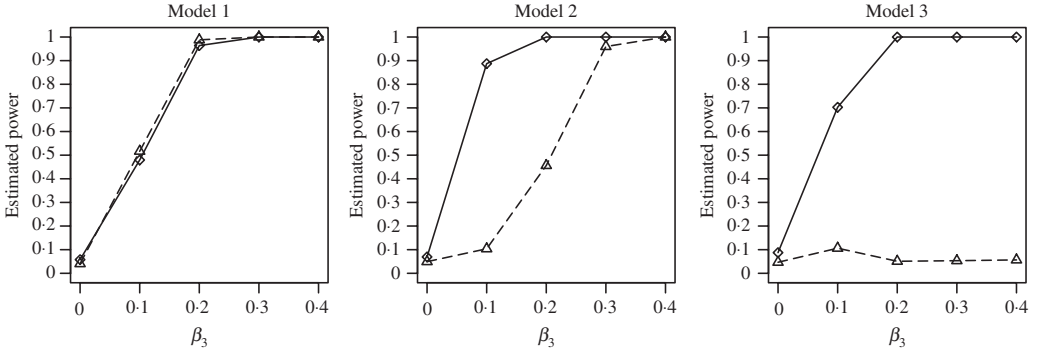

Fig. 1. Estimated power curves under Dorfman decoding. The solid and dashed curves correspond to our approach and the parametric techniques, respectively.

$\delta = 0{\cdot}1$. The analogous table and figure for master pool testing are provided in the Supplementary Material. Under both group testing algorithms, summaries of the parameter estimates pertaining to other considered values of δ were practically identical and power curves constructed for the other values of c resulted in the same conclusions. Consequently, these additional results were omitted for brevity.

Table 2. *Summary of results for data arising from Dorfman decoding*

	Parameter	Measure	$c = 1$	$c = 2$	$c = 5$	$c = 10$
Chlamydia	β_{01}	MEAN (SE)	81·7 (6·3)	82·9 (6·4)	82·7 (6·1)	82·6 (6·2)
	β_{02}	MEAN (SE)	−41·3 (9·2)	−40·4 (9·5)	−41·4 (9·1)	−39·9 (9·3)
	β_{03}	MEAN (SE)	38·8 (14·7)	37·7 (15·3)	36·8 (14·8)	37·9 (14·7)
	Percentage reduction in testing			34·0%	45·7%	35·4%
Gonorrhea	β_{01}	MEAN (SE)	47·6 (5·1)	47·7 (2·4)	48·1 (2·5)	47·1 (2·8)
	β_{02}	MEAN (SE)	−70·0 (7·8)	−69·8 (3·6)	−70·0 (3·8)	−71·2 (4·3)
	β_{03}	MEAN (SE)	50·4 (11·3)	53·1 (5·7)	52·5 (5·8)	51·3 (6·3)
	Percentage reduction in testing			45·9%	71·0%	74·0%

MEAN, mean ($\times 100$) of the 500 estimates; SE, average standard error ($\times 100$).

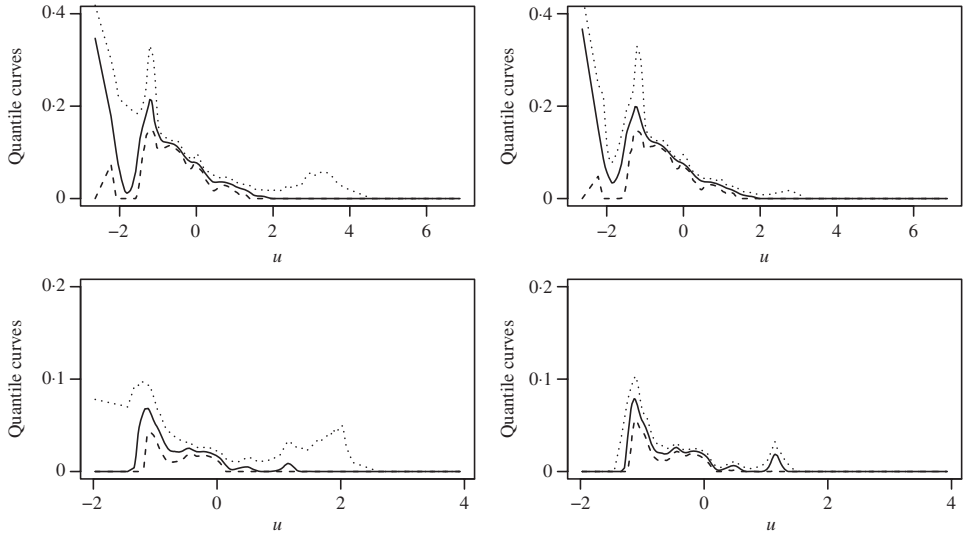

Fig. 2. Pointwise quantile curves as a function of the linear predictor u. Top row: chlamydia data, bottom row: gonorrhea data. Left column: $c = 1$, right column: $c = 5$. The dashed, solid, and dotted lines correspond to the 0·025, 0·5, and 0·975 quantiles, respectively.

5. APPLICATION TO CHLAMYDIA AND GONORRHEA DATA

In this section we illustrate our methodology using chlamydia and gonorrhea data collected by the Nebraska Public Health Laboratory. This laboratory tests patients individually for the presence of these bacterial infections, whereas other such laboratories have adopted group testing strategies; e.g., the Iowa Hygienic Laboratory uses a Dorfman type algorithm (Jirsa, 2008) to screen for these sexually transmitted diseases. The data we consider consist of individual-level testing responses obtained from assaying urine specimens collected from $N = 7310$ female patients. In addition to these testing responses we also have access to several predictor variables: namely, X_1, standardized age; X_2, a binary variable indicating the presence of symptoms, with 1 indicating symptoms were present; and X_3, a binary variable indicating the purpose of screening, with 1 indicating family planning. Using these data, we are able to artificially construct group testing data, treating the testing responses available in the dataset as the individuals' true infection statuses. We then assigned each of the individuals to a group of size c based on their specimen arrival date, where $c \in \{1, 2, 5, 10\}$. Dorfman decoding was implemented to screen the groups for both diseases, where testing responses for chlamydia and gonorrhea were simulated using the

sensitivities 0·947 and 0·913 and specificities 0·989 and 0·993, respectively. These specifications were chosen to emulate the protocol and assay currently used by the Iowa Hygienic Laboratory. This process was repeated 500 times for each value of c and our model was fit to each resulting dataset.

Table 2 provides a summary of the parameter estimates obtained from analysing the Dorfman decoding data. The regression parameter estimates obtained by our methodology are similar across all values of c, and in many situations exhibit less variability than the estimates based on the artificial individual-level data; i.e., when $c = 1$. Figure 2 provides 0·025, 0·5, and 0·975 pointwise quantile curves of the 500 estimated regression functions obtained from analysing the Dorfman decoding data when $c = 1$ and 5. The analogous figures for $c = 2$ and 10 are provided in the Supplementary Material. The estimated regression curves based on the group testing data exhibit less variability when compared to those based on individual screening data. These results indicate that through group testing the screening cost for chlamydia and gonorrhea can be reduced by up to 45·7% and 74·0%, respectively, while providing more precise inference.

Acknowledgement

We are grateful to the editor, associate editor, and referees for their helpful suggestions. We also thank Dr Joshua Tebbs for his insightful comments.

Supplementary material

Supplementary material available at *Biometrika* online provides the details of our methodology mentioned in § 2 and § 3, as well as a proof of Theorem 1 and additional simulation results. Code, written in R, that implements our new techniques, is available upon request.

Appendix

We now provide regularity conditions under which Theorem 1 in § 3 holds.

Condition A1. The functions $d_\beta(u) = E(X \mid X^{\mathrm{T}}\beta = u)$ and $p_\beta(u)$ have bounded and continuous second order derivatives.

Condition A2. The density function of $X^{\mathrm{T}}\beta$ is bounded away from zero and satisfies a Lipschitz condition of order 1 on $\{u = x^{\mathrm{T}}\beta : x \in \mathbb{X}\}$.

Condition A3. The bandwidth $h = CN^{-1/5}$ for some constant $C > 0$, and $K(\cdot)$ is a bounded and symmetric density function with bounded first derivative.

Condition A4. The function $M(\cdot, \cdot, \cdot)$ is bounded away from 0.

Condition A5. The equation $\beta^{\mathrm{T}}\Omega\beta = 0$ has the unique root $\beta = \beta_0$ in $\mathcal{B}$.

Conditions A1–A3 are common in the single-index literature. The Lipschitz condition in Condition A2 allows for discrete predictor variables. Condition A4 is easily satisfied when the assay is imperfect, as long as $0{\cdot}5 < S_e, S_p < 1$. This also assures that the denominator in Ω is bounded away from 0. Condition A5 guarantees that the matrix $\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0$ is positive definite.

REFERENCES

BILDER, C. R. & TEBBS, J. M. (2005). Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs. *Biomet. J.* **47**, 502–16.

BILDER, C. R. & TEBBS, J. M. (2009). Bias, efficiency and agreement for group-testing regression models. *J. Statist. Comp. Simul.* **79**, 67–80.

CHEN, P., TEBBS, J. M. & BILDER, C. R. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65**, 1270–8.

CUI, X., HÄRDLE, W. & ZHU, L. (2011). The EFM approach for single-index models. *Ann. Statist.* **39**, 1658–88.

DELAIGLE, A. & HALL, P. (2012). Nonparametric regression with homogeneous group testing data. *Ann. Statist.* **40**, 131–58.

DELAIGLE, A. & MEISTER, A. (2011). Nonparametric regression analysis for group testing data. *J. Am. Statist. Assoc.* **106**, 640–50.

DORFMAN, R. (1943). The detection of defective members of large populations. *Ann. Math. Statist.* **14**, 436–40.

GASTWIRTH, J. (2000). The efficiency of pooling in the detection of rare mutations. *Am. J. Hum. Genet.* **67**, 1036–9.

HÄRDLE, W., HALL, P. & ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157–8.

HUANG, X. & TEBBS, J. M. (2009). On latent-variable model misspecification in structural measurement error models for binary response. *Biometrics* **65**, 710–8.

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *J. Economet.* **58**, 71–120.

JIRSA, S. (2008). Pooling specimens: A decade of successful cost savings. National STD Prevention Conference, 2008. Chicago, IL.

KIM, H., HUDGENS, M., DREYFUSS, J., WESTREICH, D. & PILCHER, C. (2007). Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*, **63**, 1152–63.

KLEIN, R. W. & SPADY, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387–421.

LEWIS, J. L., LOCKARY, V. M. & KOBIC, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. *Sex. Transm. Dis.* **39**, 46–8.

LIN, W. & KULASEKERA, K. B. (2007). Identifiability of single-index models and additive-index models. *Biometrika* **94**, 496–501.

MACK, Y. P. & SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. verw. Geb.* **61**, 405–15.

MUÑOZ-ZANZI, C., JOHNSON, W., THURMOND, M. & HIETALA, S. (2000). Pooled-sample testing as a herd-screening tool for detection of bovine viral diarrhea virus persistently infected cattle. *J. Vet. Diagnos. Invest.* **12**, 195–203.

REMLINGER, K., HUGHES-OLIVER, J., YOUNG, S. & LAM, R. (2006). Statistical design of pools using optimal coverage and minimal collision. *Technometrics* **48**, 133–43.

VAN, T., MILLER, J., WARSHAUER, D., REISDORF, E., JERRIGAN, D., HUMES, R. & SHULT, P. (2012). Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR. *J. Clin. Microbiol.* **50**, 891–6.

VANSTEELANDT, E., GOETGHEBEUR, E. & VERSTRAETEN, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56**, 1126–33.

VENETTE, R., MOON, R. & HUTCHINSON, W. (2002). Strategies and statistics of sampling for rare individuals. *Ann. Rev. Entomol.* **47**, 143–74.

WANG, J., XUE, L., ZHU, L. & CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* **38**, 246–74.

WANG, D., ZHOU, H. & KULASEKERA, K. B. (2013). A semi-local likelihood regression estimator of the proportion based on group testing data. *J. Nonparam. Statist.* **25**, 209–21.

XIA, C. (2006). Asymptotic distributions for two estimators of the single-index model. *Economet. Theory* **22**, 1112–37.

XIA, Y., TONG, H., LI, W. K. & ZHU, L. (2002). An adaptive estimation of dimension reduction space. *J. R. Statist. Soc.* B **64**, 363–410.

XIE, M. (2001). Regression analysis of group testing samples. *Statist. Med.* **20**, 1957–69.

ZHANG, B., BILDER, C. & TEBBS, J. (2013). Group testing regression model estimation when case identification is a goal. *Biomet. J.* **55**, 173–89.

ZHU, L. & XUE, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J. R. Statist. Soc.* B **68**, 549–70.

[*Received May* 2013. *Revised January* 2014]