Published in final edited form as:

Stat Med. 2017 December 30; 36(30): 4860–4872. doi:10.1002/sim.7455.

Group testing regression models with dilution submodels

Md S. Warasi^a, Christopher S. McMahan^b, Joshua M. Tebbs^{c,*}, and Christopher R. Bilder^d

^aDepartment of Mathematics and Statistics, Radford University, Radford, VA 24142, USA

^bDepartment of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA

^cDepartment of Statistics, University of South Carolina, Columbia, SC 29208, USA

^dDepartment of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

Abstract

Group testing, where specimens are tested initially in pools, is widely used to screen individuals for sexually transmitted diseases. However, a common problem encountered in practice is that group testing can increase the number of false negative test results. This occurs primarily when positive individual specimens within a pool are diluted by negative ones, resulting in positive pools testing negatively. If the goal is to estimate a population-level regression model relating individual disease status to observed covariates, severe bias can result if an adjustment for dilution is not made. Recognizing this as a critical issue, recent binary regression approaches in group testing have utilized continuous biomarker information to acknowledge the effect of dilution. In this paper, we have the same overall goal but take a different approach. We augment existing group testing regression models (that assume no dilution) with a parametric dilution submodel for pool-level sensitivity and estimate all parameters using maximum likelihood. An advantage of our approach is that it does not rely on external biomarker test data, which may not be available in surveillance studies. Furthermore, unlike previous approaches, our framework allows one to formally test whether dilution is present based on the observed group testing data. We use simulation to illustrate the performance of our estimation and inference methods, and we apply these methods to two infectious disease data sets.

Keywords

Binary regression; dilution effect; likelihood ratio test; maximum likelihood; pooled testing; sensitivity

1. Introduction

Group testing, which is also known as pooled testing, involves combining individual specimens (e.g., blood, plasma, swabs, urine, etc.) into a pool and then testing the pool for a binary characteristic, such as disease status. Dorfman [1] is usually credited with its introduction, having proposed it to screen United States soldiers for syphilis during World

*Correspondence to: 209G LeConte, Department of Statistics, University of South Carolina, Columbia, SC 29208, USA, tebbs@stat.sc.edu.

War II. Now widely regarded for its cost-savings potential, group testing is commonly used in large-scale sexually transmitted disease screening applications involving HIV [2], HBV/HCV [3, 4], chlamydia and gonorrhea [5], and for other infections including West Nile virus [6] and influenza [7]. Pooling individuals arises more generally in environmental monitoring [8], entomology [9], genetics [10], animal testing [11], drug discovery [12], and food safety [13].

Statistical research in group testing has traditionally split into two areas: case identification and estimation. Case identification involves classifying each individual subject as diseased or not [14, 15], whereas estimation involves estimating a population-level prevalence [16, 17] or a population-level regression model relating individual disease status to covariates [18, 19]. When the disease prevalence is small, group testing can drastically reduce the number of tests needed in the case identification problem. Furthermore, estimates of the overall prevalence and estimates of population-level regression parameters can be more precise than if one were to test subjects individually [17, 20].

A widespread concern attached to group testing is the so-called “dilution effect,” which refers to a positive pool testing negatively because an assay cannot detect positive material within the pool. This situation arises frequently in infectious disease testing, where it is well known that assay sensitivity can decrease when the pool size is too large. Previous research in group testing has considered the dilution effect in a homogeneous population. Wein and Zenios [21] proposed a parametric hierarchical model that relates the continuous response of an assay to the latent biomarker concentration in the pool and used dynamic programming to generalize Dorfman’s two-stage testing algorithm (for case identification) while minimizing costs and the number of false negative results. In their companion paper, Zenios and Wein [22] used this same modeling framework to improve estimation of a population-level prevalence. Other works include Hwang [23], who determined optimal Dorfman pool sizes in the presence of dilution (for a specific dilution model) and Hung and Swallow [24], who examined robustness properties for estimation. In the regression problem, McMahan et al. [25], Wang et al. [26], and Delaigle and Hall [27] have each generalized the biomarker distribution framework proposed in [21, 22] to include covariate information. McMahan et al. [25] and Wang et al. [26] accomplish this by using parametric methods, while Delaigle and Hall [27] develop nonparametric estimators.

The conclusion reached by those authors who have examined the dilution effect is that severe bias in parameter estimation can arise if dilution is present and not acknowledged. In the regression problem, this bias can potentially lead to making incorrect conclusions about the covariate effects and also to subject-specific probabilities of disease that are severely underestimated. In this paper, we take a different approach to account for dilution with group testing data in a regression setting. Specifically, we elicit a parametric submodel for pool-level sensitivity, similar to what Hung and Swallow [24] and Zhang et al. [28] used in considering estimation within a homogeneous population, and then estimate the population-level regression model and dilution submodel jointly. This approach offers two distinct advantages when compared to competing methods. First, unlike [25, 26, 27], we do not require information about the latent biomarker distributions associated with an assay or training data to estimate them. These distributions are rarely known in surveillance studies,

for example, whereas one can easily construct a dilution submodel using any cumulative distribution function (see Section 2). Second, using the information in the observed covariates and group testing responses, one can within our modeling framework perform a hypothesis test to detect whether dilution is present (see Section 4). Having such knowledge (i.e., whether or not a dilution effect exists) can be invaluable to investigators who are designing future studies and to those who are considering whether using group testing is feasible.

Subsequent sections of this article are organized as follows. In Section 2, we introduce notation, state the assumptions regarding our use of a dilution submodel, and give examples of submodels that can be used in practice. In Section 3, we describe our regression estimation techniques for two group testing scenarios: master pool testing (for estimation purposes only) and Dorfman testing (for estimation and case identification). In Section 4, we present a likelihood ratio test to detect dilution with the observed group testing data. In Section 5, we use simulation to assess the performance of our methods, and, in Section 6, we apply them to two infectious disease data sets. In Section 7, we conclude with a summary and discuss future research. We use the Web-based Supporting information to provide derivations and additional simulation results.

2. Preliminaries

2.1. Notation and assumptions

Suppose each of N individuals is assigned to exactly one non-overlapping master pool. Let $\tilde{Y}_{ij} = 1$ if the i th individual in the j th pool is truly positive, $\tilde{Y}_{ij} = 0$ otherwise, for $i = 1, 2, \dots, c_j$ and $j = 1, 2, \dots, J$, so that $N = \sum_{j=1}^J c_j$. We call c_j the pool size for the j th pool and J denotes the number of master pools tested. Let $\mathbf{x}_{ij} = (1, x_{ij1}, \dots, x_{ijr})'$ denote an $(r+1) \times 1$ vector of covariates observed for the i th individual in the j th pool. Throughout this article, we assume the $\tilde{Y}_{ij}$'s are conditionally independent given the covariates with

$$p_{ij} = \text{pr}(\tilde{Y}_{ij} = 1 | \mathbf{x}_{ij}) = g^{-1}(\mathbf{x}_{ij}'\boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_r)'$ is a vector of regression coefficients and $g(\cdot)$ is a monotone and differentiable link function. We also assume throughout that the covariates $\mathbf{x}_{ij}$ are best regarded as fixed (i.e., not random). Let $\tilde{Z}_j = 1$ if the j th pool contains at least one positive individual, $\tilde{Z}_j = 0$ otherwise, so that

$$\text{pr}(\tilde{Z}_j = 1) = 1 - \prod_{i=1}^{c_j} (1 - p_{ij}) = 1 - \prod_{i=1}^{c_j} \{1 - g^{-1}(\mathbf{x}_{ij}'\boldsymbol{\beta})\}.$$

Note that because of inherent assay error, the true master pool statuses $\tilde{Z}_j$ and the true individual statuses $\tilde{Y}_{ij}$ are best regarded as latent.

To perform a regression analysis, one is left to estimate the model

$p_{ij} = \text{pr}(\tilde{Y}_{ij}=1 | x_{ij}) = g^{-1}(x'_{ij}\beta)$ with the observed group testing responses. Let $Z_j = 1$ if the j th master pool tests positively, $Z_j = 0$ otherwise, for $j = 1, 2, \dots, J$. Previous research in group testing regression; see, e.g., [18, 20], has assumed that the sensitivity of the assay used to test master pools does not depend on the size of the pool; i.e.,

$$S_e = \text{pr}(Z_j=1 | \tilde{Z}_j=1) = \text{pr}(Z_{j'}=1 | \tilde{Z}_{j'}=1), \quad (1)$$

for all j, j' , even if $c_j = c_{j'}$. This assumption also implies $\text{pr}(Z_j=1 | \tilde{Z}_j=1) = S_e$ for individual testing; i.e., a master pool of size $c_j = 1$. From Bayes' theorem and the Law of Total Probability, it is insightful to note that

$$S_e = \text{pr}(Z_j=1 | \tilde{Z}_j=1) = \frac{\sum_{k=1}^{c_j} \text{pr}(Z_j=1 | \sum_{i=1}^{c_j} \tilde{Y}_{ij}=k) \text{pr}(\sum_{i=1}^{c_j} \tilde{Y}_{ij}=k)}{\text{pr}(\tilde{Z}_j=1)}.$$

Therefore, for the constant sensitivity assumption in (1) to hold, one would have to require that the probabilities $\text{pr}(Z_j=1 | \sum_{i=1}^{c_j} \tilde{Y}_{ij}=k)$ are equal for each $k = 1, 2, \dots, c_j$. Of course, such a requirement is unsuitable if testing results are influenced by dilution. As noted earlier, McMahan *et al.* [25], Wang *et al.* [26], and Delaigle and Hall [27] relax this assumption by using the underlying (latent) biomarker distributions associated with an assay to calculate $\text{pr}(Z_j=1 | \sum_{i=1}^{c_j} \tilde{Y}_{ij}=k)$ for each k ; see, e.g., Section 3.2 in [25]. In this article, we incorporate dilution differently.

2.2. Dilution submodels

To account for dilution in the group testing regression problem without making assumptions about the biomarker distributions, we specify, for $\lambda \geq 0$,

$$h(k, c_j, \lambda) = \text{pr}_\lambda \left(Z_j=1 | \sum_{i=1}^{c_j} \tilde{Y}_{ij}=k \right),$$

as a submodel for the pool-level sensitivity, where the function $h(\cdot, \cdot, \cdot)$ satisfies the following two conditions:

- i. $h(k, c_j, \lambda) \leq h(k', c_j, \lambda)$, for all $k < k'$
- ii. $h(k, c_j, \lambda) \leq h(k, c_{j'}, \lambda)$, for all $c_j > c_{j'}$.

The first condition says that for equally sized pools, the sensitivity is monotonically increasing in k , the number of positive individuals. The second condition says that for pools with the same number of positive individuals, the sensitivity is monotonically decreasing in the pool size. These two characteristics describe mathematically what is commonly referred to as the “dilution effect” in the group testing literature. Henceforth, we call λ the “dilution

parameter.” Our estimation methods described in Section 3 are aimed at estimating the regression parameter β and λ jointly.

How should one specify a parametric form for the dilution submodel $h(k, c_j, \lambda)$? When one couples the monotonicity requirements in (i) and (ii) with the fact that the sensitivity is itself a probability, it is natural to consider cumulative distribution functions as a basis for its specification. We proceed under the following mild assumptions:

(A1) The assay sensitivity for individual testing is S_e , where S_e is known.

(A2) A pool comprised of all positive individuals has sensitivity equal to S_e ; i.e.,

$$\text{pr}_{\lambda}(Z_j=1|\sum_{i=1}^{c_j} \tilde{Y}_{ij}=c_j)=S_e.$$

Neither of these assumptions is prohibitive. Before an assay is released for general use, manufacturers conduct pilot studies with known positive individuals to determine (or at least estimate) S_e for individual testing. The second assumption is reasonable because if a pool consists entirely of positive individuals, there is no possibility of dilution.

As an example, consider the logistic cumulative distribution function and assume that

$$\text{logit}\{h(k, c_j, \lambda)\}=\lambda_0+\lambda(k/c_j), \quad (2)$$

where $\lambda_0 \in \mathbb{R}$ and $\lambda \geq 0$ are both constants. Clearly, the monotonicity requirements in (i) and (ii) hold, and assumptions (A1) and (A2) require that λ_0 be chosen to satisfy $h(k, c_j, \lambda) = S_e$ when $k = c_j$; i.e.,

$$\frac{\exp(\lambda_0+\lambda)}{1+\exp(\lambda_0+\lambda)}=S_e \Rightarrow \lambda_0=\ln\left(\frac{S_e}{1-S_e}\right)-\lambda.$$

Substituting this value into (2) and simplifying, we obtain

$$h(k, c_j, \lambda)=\frac{\exp\{\lambda\tau(k, c_j)\}}{(1/S_e)+\exp\{\lambda\tau(k, c_j)\}-1}, \quad (3)$$

where $\tau(k, c_j) = (k - c_j)/c_j$. We call this the logistic dilution submodel.

For illustration, Table 1 shows numerical values of the logistic dilution submodel $h(k, c_j, \lambda)$ when the dilution parameter $\lambda \in \{2.6, 3.8, 5.0\}$, $S_e = 0.99$, and $c_j \in \{5, 10\}$, for $k = 1, 2, \dots, c_j$. This table is instructive because it allows one to verify the monotonicity requirements in (i) and (ii) for a specific dilution submodel. Values of λ smaller (larger) than those shown in Table 1 correspond to smaller (larger) amounts of dilution, and $h(k, c_j, \lambda) = S_e$ when $\lambda = 0$; i.e., there is no dilution. By using similar arguments, it is easy to develop other submodels; for example, with the standard normal and Gumbel cumulative distribution functions, respectively, one can show that

$$h(k, c_j, \lambda) = \Phi\{\Phi^{-1}(S_e) + \lambda\tau(k, c_j)\},$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution, and

$$h(k, c_j, \lambda) = 1 - \exp(-\exp[\ln\{-\ln(1 - S_e)\} + \lambda\tau(k, c_j)])$$

also satisfy (i) and (ii). We call these the probit and the complementary log-log dilution submodels, respectively. Tables for these dilution submodels (similar to Table 1) and a different submodel considered by Hung and Swallow [24] are shown in the Web-based Supporting information.

3. Estimation

Our goal is to estimate the binary regression model $p_{ij} = \text{pr}(\tilde{Y}_{ij} = 1 | \mathbf{x}_{ij}) = g^{-1}(\mathbf{x}_{ij}'\boldsymbol{\beta})$ while incorporating dilution through the submodel $h(k, c_j, \lambda)$. We first consider estimation by utilizing only the master pool responses $\mathbf{Z} = (Z_1, Z_2, \dots, Z_J)$, similar to what McMahan *et al.* [25] and Delaigle and Hall [27] considered within their biomarker framework. We then extend this to group testing data that arise from Dorfman testing, which is the most widely used case identification algorithm in human disease testing applications and elsewhere.

3.1. Master pool testing

The data available for modeling with master pool testing (MPT) is $\mathcal{D}_M = \{\mathbf{Z}, \mathbf{X}\}$, where $\mathbf{Z}$ is the collection of master pool responses and $\mathbf{X}$ is the collection of covariates measured for each of the N individuals. We make a third assumption for the remainder of this article:

- (A3) The assay specificity; i.e., $\text{pr}(Z_j = 0 | \tilde{Z}_j = 0) = S_p$, where S_p is known and does not depend on c_j .

Assumption (A3) is very common in the group testing literature; see, e.g., [14], and is reasonable in practice because the event $\{\tilde{Z}_j = 0\}$ can only occur if all contributing individuals are truly negative; i.e., the specimen tested is completely disease free. From the Law of Total Probability, we can write

$$\text{pr}(Z_j = 1) = (1 - S_p) \prod_{i=1}^{c_j} (1 - p_{ij}) + \sum_{k=1}^{c_j} h(k, c_j, \lambda) \text{pr}\left(\sum_{i=1}^{c_j} \tilde{Y}_{ij} = k\right).$$

Finding $\text{pr}(\sum_{i=1}^{c_j} \tilde{Y}_{ij} = k)$ involves working with the sum of independent but non-identically distributed Bernoulli random variables and can be calculated using the approach in [29]. The observed data likelihood from MPT is given by

$$L(\boldsymbol{\theta}|\mathcal{D}_M) = \prod_{j=1}^J \{\text{pr}(Z_j=1)\}^{Z_j} \{1 - \text{pr}(Z_j=1)\}^{1-Z_j}, \quad (4)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \lambda)'$. The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\lambda})'$, can be found by maximizing $L(\boldsymbol{\theta}|\mathcal{D}_M)$ directly. Towards calculating (large-sample) standard errors, the negative inverse Hessian of the logarithm of (4) evaluated at $\hat{\boldsymbol{\theta}}$ can be used to approximate the covariance matrix of $\hat{\boldsymbol{\theta}}$.

3.2. Dorfman testing

Dorfman testing (DT) is a two-stage hierarchical procedure where non-overlapping pools are tested in the first stage (like MPT). Pools that test positively are then resolved by retesting each individual separately in the second stage. Intuitively, if additional retesting responses on positive pools are available, including them in the analysis should sharpen one's inference for the covariate effects. In fact, in the absence of dilution, Zhang *et al.* [20] found that regression estimates from DT were even more precise than the corresponding estimates from individual testing-despite requiring fewer tests. Wang *et al.* [26] observed the same phenomenon in the presence of dilution from within their biomarker framework.

The observed data from DT is $\mathcal{D}_D = \{\mathbf{Z}, \mathbf{Y}, \mathbf{X}\}$, where $\mathbf{Z}$ contains the master pool responses (as before), $\mathbf{Y}$ is the collection of stage-2 individual testing responses, and $\mathbf{X}$ aggregates the individual covariates (as before). To accommodate data of this structure for a regression analysis, we assume that the observed testing responses in $\mathbf{Z}$ and $\mathbf{Y}$ are independent, conditional on the true statuses of the individuals, which we denote by $\tilde{\mathbf{Y}}$. This assumption is common in the group testing literature when case identification is the ultimate goal [14] and is practicable when testing errors arise by factors related to test implementation. Under this assumption, the observed data likelihood can be expressed as

$$L(\boldsymbol{\theta}|\mathcal{D}_D) = \sum_{\tilde{\mathbf{Y}} \in \{0,1\}^N} T_1(\boldsymbol{\beta}, \mathcal{D}_D, \tilde{\mathbf{Y}}) T_2(\lambda, \mathcal{D}_D, \tilde{\mathbf{Y}}) T_3(\mathcal{D}_D, \tilde{\mathbf{Y}}), \quad (5)$$

where

$$T_1(\boldsymbol{\beta}, \mathcal{D}_D, \tilde{\mathbf{Y}}) = \prod_{j=1}^J \prod_{i=1}^{c_j} p_{ij}^{\tilde{Y}_{ij}} (1 - p_{ij})^{1-\tilde{Y}_{ij}}$$

$$T_2(\lambda, \mathcal{D}_D, \tilde{\mathbf{Y}}) = \prod_{j=1}^J \prod_{k=1}^{c_j} [h(k, c_j, \lambda)^{Z_j} \{1 - h(k, c_j, \lambda)\}^{1-Z_j}]^{I_{jk}}$$

$$T_3(\mathcal{D}_D, \tilde{\mathbf{Y}}) = \prod_{j=1}^J \{S_p^{1-Z_j} (1 - S_p)^{Z_j}\}^{I_{j0}} \left[\prod_{i=1}^{c_j} \{S_e^{\tilde{Y}_{ij}} (1 - S_p)^{1-\tilde{Y}_{ij}}\}^{Y_{ij}} \{S_p^{1-\tilde{Y}_{ij}} (1 - S_e)^{\tilde{Y}_{ij}}\}^{1-Y_{ij}} \right]^{Z_j},$$

where $I_{jk} = \mathcal{I}(\sum_{i=1}^{c_j} \tilde{Y}_{ij} = k)$, for $k = 0, 1, \dots, c_j$ and $\mathcal{I}(\cdot)$ is the indicator function. Note that in $T_3(\mathcal{D}_D, \tilde{\mathbf{Y}})$, we denote by Y_{ij} the binary testing response for the i th individual in the j th pool, if this individual is retested separately, which only occurs when $Z_j = 1$.

Because the sum in (5) is taken over all 2^N possible true individual statuses in $\tilde{\mathbf{Y}}$, evaluating $L(\boldsymbol{\theta} | \mathcal{D}_D)$ directly can be computationally burdensome. To circumvent this complexity, we develop an expectation-maximization (EM) algorithm to find the MLE of $\boldsymbol{\theta}$ by viewing the individuals' true statuses as "missing data." The complete data likelihood can be expressed as

$$L_C(\boldsymbol{\theta} | \mathcal{D}_D, \tilde{\mathbf{Y}}) = T_1(\boldsymbol{\beta}, \mathcal{D}_D, \tilde{\mathbf{Y}}) T_2(\lambda, \mathcal{D}_D, \tilde{\mathbf{Y}}) T_3(\mathcal{D}_D, \tilde{\mathbf{Y}}). \quad (6)$$

The E-step finds $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) = E\{L_C(\boldsymbol{\theta} | \mathcal{D}_D, \tilde{\mathbf{Y}}) | \mathcal{D}_D, \boldsymbol{\theta}^{(d)}\}$, where the expectation is taken with respect to the individuals' true (latent) statuses and $L_C(\boldsymbol{\theta} | \mathcal{D}_D, \tilde{\mathbf{Y}})$ denotes the natural logarithm of (6). Note that $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ can be expressed as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)}) = \mathcal{T}_1(\boldsymbol{\beta}, \mathcal{D}_D, \boldsymbol{\theta}^{(d)}) + \mathcal{T}_2(\lambda, \mathcal{D}_D, \boldsymbol{\theta}^{(d)}) + \mathcal{T}_3(\mathcal{D}_D, \boldsymbol{\theta}^{(d)}),$$

where

$$\mathcal{T}_1(\boldsymbol{\beta}, \mathcal{D}_D, \boldsymbol{\theta}^{(d)}) = \sum_{j=1}^J \sum_{i=1}^{c_j} E(\tilde{Y}_{ij} | \mathcal{D}_D, \boldsymbol{\theta}^{(d)}) \ln p_{ij} + \{1 - E(\tilde{Y}_{ij} | \mathcal{D}_D, \boldsymbol{\theta}^{(d)})\} \ln(1 - p_{ij}),$$

$$\mathcal{T}_2(\lambda, \mathcal{D}_D, \boldsymbol{\theta}^{(d)}) = \sum_{j=1}^J \sum_{k=1}^{c_j} [Z_j \ln h(k, c_j, \lambda) + (1 - Z_j) \ln \{1 - h(k, c_j, \lambda)\}] E(I_{jk} | \mathcal{D}_D, \boldsymbol{\theta}^{(d)}),$$

and $\mathcal{T}_3(\mathcal{D}_D, \boldsymbol{\theta}^{(d)})$ is free of $\boldsymbol{\theta}$. We provide closed-form expressions for $E(\tilde{Y}_{ij} | \mathcal{D}_D, \boldsymbol{\theta}^{(d)})$ and $E(I_{jk} | \mathcal{D}_D, \boldsymbol{\theta}^{(d)})$ in the Web-based Supporting information. The M-step then finds $\boldsymbol{\theta}^{(d+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$. Given the form of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(d)})$ above, the maximization step can be divided into two stages; i.e., one can find $\boldsymbol{\theta}^{(d+1)} = (\boldsymbol{\beta}^{(d+1)})', \lambda^{(d+1)}'$, where

$$\boldsymbol{\beta}^{(d+1)} = \arg \max_{\boldsymbol{\beta}} \mathcal{T}_1(\boldsymbol{\beta}, \mathcal{D}_D, \boldsymbol{\theta}^{(d)})$$

and then

$$\lambda^{(d+1)} = \arg \max_{\lambda \geq 0} \mathcal{F}_2(\lambda, \mathcal{D}, \boldsymbol{\theta}^{(d)}).$$

The optimization step to obtain $\boldsymbol{\beta}^{(d+1)}$ can use any standard optimization routine for fitting binary regression models; e.g., `optim` in R. Similarly, because $\lambda \geq 0$, finding $\lambda^{(d+1)}$ requires a search over a one-dimensional constrained space; this search can use standard numerical optimization routines; e.g., `optimize` in R.

We now summarize the EM algorithm to estimate $p_{ij} = \text{pr}(\tilde{Y}_{ij} = 1 | \mathbf{x}_{ij}) = g^{-1}(\mathbf{x}_{ij}' \boldsymbol{\beta})$ while incorporating dilution through $h(k, c_j; \lambda)$ for DT. Initialize $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)'}, \lambda^{(0)'})'$, set $d = 0$, and repeat the following steps until convergence:

1. (E-Step) Evaluate $E(I_{jk} | \mathcal{D}_D, \boldsymbol{\theta}^{(d)})$, for $k = 1, 2, \dots, c_j$ and $j = 1, 2, \dots, J$.
2. (E-Step) Evaluate $E(\tilde{Y}_{ij} | \mathcal{D}_D, \boldsymbol{\theta}^{(d)})$, for $i = 1, 2, \dots, c_j$ and $j = 1, 2, \dots, J$.
3. (M-Step) Obtain $\boldsymbol{\theta}^{(d+1)} = (\boldsymbol{\beta}^{(d+1)'}, \lambda^{(d+1)'})'$, as described in the last paragraph.
4. Set $d = d + 1$ and return to Step 1.

At convergence, take $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(d)}$ to be the MLE of $\boldsymbol{\theta}$. One can find the observed information matrix ($\boldsymbol{\theta}$) by appealing to the missing data principle and the method outlined in [30]; i.e.,

$$(\boldsymbol{\theta}) = - \frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} - \text{cov} \left\{ \frac{\partial l_G(\boldsymbol{\theta} | \mathcal{D}_D, \tilde{\mathbf{Y}})}{\partial \boldsymbol{\theta}} \middle| \mathcal{D}_D, \boldsymbol{\theta} \right\}. \quad (7)$$

Expressions for both terms on the right-hand side of (7) are given in the Web-based Supporting information. Large-sample Wald inference can be conducted in the usual way by using $(\hat{\boldsymbol{\theta}})^{-1}$ as an estimate of $(\boldsymbol{\theta})^{-1}$.

4. Likelihood ratio test for dilution

An advantage of our approach to incorporate dilution in the group testing regression problem is that one can formally test for dilution. To see how, consider any submodel $h(k, c_j; \lambda)$ as described in Section 2.2 (e.g., the logistic, probit, and complementary log-log submodels). Under assumptions (A1) and (A2), note that when $\lambda = 0$ one obtains $h(k, c_j; 0) = S_e$ for each k and c_j ; i.e., there is no dilution. This naturally suggests the following test

$$H_0: \lambda = 0 \quad \text{versus} \quad H_1: \lambda > 0$$

as a way to infer whether dilution is present. Of course, if the observed data are consistent with H_0 , then simpler modeling techniques can be used to estimate

$$p_{ij} = \text{pr}(\tilde{Y}_{ij} = 1 | \mathbf{x}_{ij}) = g^{-1}(\mathbf{x}_{ij}' \boldsymbol{\beta}) \text{ under MPT [18] or DT [19, 20].}$$

Because we are using likelihood-based estimation methods, large-sample Wald, score, and likelihood ratio procedures all emerge as possible testing strategies. Tests like H_0 versus H_1 have been described as “nonstandard” in the constrained inference literature [31], because the value of λ under H_0 corresponds to a boundary point of the parameter space. Based on the recommendations in Molenberghs and Verbeke [32], we focus on the likelihood ratio test (LRT). In our testing problem, the LRT statistic is given by

$$T_{LR} = -2 \ln \left\{ \frac{\sup_{\theta \in \Theta_0} L(\theta | \mathcal{D})}{\sup_{\theta \in \Theta} L(\theta | \mathcal{D})} \right\},$$

where $\Theta = \{(\beta', \lambda) : \beta \in \mathbb{R}^{r+1}, \lambda \geq 0\}$ denotes the full parameter space, $\Theta_0 = \{(\beta', \lambda) : \beta \in \mathbb{R}^{r+1}, \lambda = 0\}$ is the reduced model parameter space, and $\mathcal{D}$ denotes the observed data (i.e., either $\mathcal{D}_M$ for MPT or $\mathcal{D}_D$ for DT). The asymptotic distribution of T_{LR} is the two-component mixture $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, where χ_0^2 is a point mass distribution at 0 and χ_1^2 denotes the central χ^2 distribution with 1 degree of freedom. Large values of T_{LR} are evidence against H_0 .

5. Simulation study

We examined the performance of our regression methods and likelihood ratio test under a variety of models which encapsulate low-prevalence situations where group testing would be feasible [33]. The findings presented herein are representative of those based on our larger collection of results.

5.1. Simulation description

We consider the binary regression model

$$\text{logit}\{\text{pr}(\tilde{Y}_{ij}=1|x_{ij})\} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}, \quad (8)$$

where $x_{ij1} \sim \mathcal{N}(0, 0.75^2)$, $x_{ij2} \sim \text{Bernoulli}(0.1)$, and $\beta = (-3, 2, 1)'$. These settings give an overall disease prevalence of about 9%, which is similar to both of our applications in Section 6. We generated $N = 5000$ individual true statuses $\tilde{Y}_{ij}$ from (8). This sample size is larger than that used in our HBV data application (in Section 6.1) but smaller than that used in the chlamydia data application (in Section 6.2). With the 5000 true individual responses, we manually create group testing data sets by

1. forming master pools randomly or homogeneously (using pool sizes $c_j = 5$ or $c_j = 10$)
2. simulating testing responses for MPT or DT; i.e., simulate $\mathbf{Z}$ for MPT or $\{\mathbf{Z}, \mathbf{Y}\}$ for DT (see Section 3).

By “homogeneous pooling,” we mean that master pools are created by combining individuals with “similar” covariate vectors x_{ij} . In the absence of dilution, Vansteelandt *et al.* [18] and Bilder and Tebbs [34] demonstrate that pooling homogeneously produces estimates

of β that are the least variable. Homogeneous pooling is performed here by first stratifying the 5000 individuals according to the binary covariate x_{ij2} and then sorting by the continuous covariate x_{ij1} . “Random pooling” means that master pools are created by random assignment of individuals, which may emulate the resulting covariate compositions when individuals are pooled chronologically [34]. For each pooling structure (random and homogeneous), there are $J = 1000$ ($J = 500$) master pools created when the pool size is $c_j = 5$ ($c_j = 10$).

To acknowledge dilution, MPT responses are simulated as $Z_j \sim \text{Bernoulli}\{h(k, c_j, \lambda)I_{jk} + (1 - S_p)I_{j0}\}$, $j = 1, 2, \dots, J$, where $h(k, c_j, \lambda)$ is the logistic dilution submodel in (3), $S_e = S_p = 0.99$, and $I_{jk} = \mathcal{I}(\sum_{i=1}^{c_j} \tilde{Y}_{ij} = k)$, for $k = 0, 1, \dots, c_j$. Consistent with Table 1, we use values of the dilution parameter $\lambda = 2.6$ (mild dilution), $\lambda = 3.8$ (moderate dilution) and $\lambda = 5.0$ (severe dilution). For DT, when necessary, individual testing responses Y_{ij} are simulated as $Y_{ij} \sim \text{Bernoulli}\{S_e \tilde{Y}_{ij} + (1 - S_p) \tilde{Y}_{ij}\}$, also with $S_e = S_p = 0.99$. This entire process was repeated 500 times, giving us $B = 500$ group testing data sets for both MPT and DT.

5.2. Simulation results

Estimation results for DT are shown in Table 2. We provide the average of the 500 maximum likelihood estimates (“Mean”), the sample standard deviation of the estimates (“SD”), the averaged standard errors (“SE”), and the empirical coverage probabilities associated with nominal 95% Wald confidence intervals for the regression parameters (“Cov”). The right-hand side of Table 2 displays results which account for dilution; see Section 3.2. The left-hand side displays the corresponding maximum likelihood results which ignore dilution; see [19, 20]. The same summaries of the dilution parameter estimates are given in the Web-based Supporting information. Estimated size and power results to detect dilution using our LRT in Section 4 are given in Table 3.

We make several observations for DT first. The right-hand side of Table 2 shows that, on average, our estimation techniques pin down the true value of β in the presence of dilution. Regardless of whether pools are constructed randomly or homogeneously, regression estimates are very close to the true values whereas the same averages which ignore dilution can be severely biased. Our approach also provides strong agreement between the sample standard deviation of the estimates (SD) and the averaged standard error (SE), suggesting that the large-sample covariance matrix of $\hat{\beta}$ is estimated correctly on average. Therefore, it is not surprising that empirical coverage probabilities of nominal 95% Wald confidence intervals are on target. This is in sharp contrast to those calculated when dilution is ignored; the left-hand side of Table 2 shows that these same probabilities are usually outside the margin of Monte Carlo sampling error (0.03, calculated at the 99% confidence level) and can be inappropriately low when the amount of dilution is larger.

Table D.1 in the Web-based Supporting information provides the same summary for MPT. Because retests are not performed with MPT, there is slightly more bias and increased variability in the maximum likelihood estimates (when compared to DT), although the amount of bias is generally small for the covariate effects—especially under lower levels of dilution. Note that, with MPT, one is essentially trying to learn about both the covariate effects and the dilution effect with fewer testing responses (1000 and 500, respectively, for

pool sizes $c_j = 5$ and $c_j = 10$) and no retesting information. Therefore, the pooling composition (random or homogeneous) under MPT can have a large impact on the resulting estimates. For example, in Table D.1, note that those estimates linked to the covariates (i.e., $\hat{\beta}_1$ and $\hat{\beta}_2$) are much more precise under homogeneous pooling than under random pooling, an intuitive and expected result based on the work in [18, 34]. We also observed this phenomenon when we repeated the MPT simulation with $N = 20000$ individuals using the same master pool sizes; see Table D.2 in the Web-based Supporting information. On the other hand, the initial pool composition under DT (see Table 2) has very little or no impact on the regression estimates. This is also expected because after positive pools are resolved, one gets to observe the follow-up individual testing responses. Whether the master pools were formed randomly or homogeneously in this situation becomes less relevant.

It is natural to wonder how much influence the chosen dilution submodel $h(k, c_j, \lambda)$ might have on the resulting regression analysis and the LRT in Section 4. Estimation and LRT size/power summaries in Tables 2, D.1, D.2, and 3 are given under the assumption that the logistic dilution submodel in (3) is correctly specified. However, for the regression estimates, we have consistently found that misspecifying the parametric form of $h(k, c_j, \lambda)$ and thus estimating β under an incorrect submodel assumption often has very little impact. This was documented by the first author in his Ph.D. dissertation [33] when using equal master pool sizes; in the Web-based Supporting information (see Tables D.3–D.6), we illustrate this by providing additional simulation results when the master pool sizes c_j are variable. When testing for dilution, Table 3 shows that our LRT confers the nominal size (i.e., when there is no dilution, $\lambda = 0$) with estimated powers for DT increasing to unity as the dilution parameter λ increases. In the Web-based Supporting information (see Table D.8 and Figures D.1–D.4), we provide additional simulation evidence that our LRT performs nominally—even when an incorrect submodel is assumed.

6. Data applications

We use two infectious disease testing data sets to illustrate our methodology. The first data set contains continuous biomarker measurements for each individual. The second data set does not. We apply DT in each application to emulate how individual diagnoses could be made and use our estimation and inference methods with the resulting data. We also provide further evidence that our estimation techniques are robust to dilution submodel misspecification.

6.1. Irish prisoner data

Allwright et al. [35] summarize the results of a national survey of prisoners in five of seven committal prisons in Ireland. In this study, the hepatitis B virus (HBV) status was determined for each participating prisoner using an enzyme immunoassay with oral fluid specimens. In addition to each prisoner's diagnosed HBV status, the continuous optical density (OD) biomarker associated with the assay was available for each testing response as well as covariate information recorded for each prisoner. These data were collected in 1998 and provided to us by Dr. Allwright in 2005. We use the $N = 1137$ individual prisoners (99

HBV positive and 1038 HBV negative cases) in the data set for whom complete biomarker and covariate information is available.

Because continuous biomarker responses are available on each individual prisoner, these data were also used by McMahan *et al.* [25], Wang *et al.* [26], and Delaigle and Hall [27] to illustrate their dilution regression methods. Similar to these articles, we create group testing data sets by pooling individuals ourselves and simulating master pool testing responses Z_j from the individual OD readings, following the same approach outlined in [25]. Proceeding in this manner allows us to illustrate two aspects of our methods. First, we can compare our regression estimates (which account for dilution) to the estimates obtained from individual testing (i.e., no dilution; based on the $N = 1137$ individuals). Second, because a continuous biomarker is used to create the master pool responses, we can evaluate our regression approach when the dilution submodel is necessarily misspecified.

We create master pools of common size $c_j \in \{2, 4, 6, 8\}$, with one smaller-sized remainder pool for each choice of c_j . Using random assignment of individual specimens to pools, we repeated this process 500 times, providing us with $B = 500$ sets of master pools for each pool size. Dorfman testing (DT) was then applied to master pools that tested positively (occurring when the average OD reading exceeded a threshold t^*), and individual diagnoses Y_{ij} were made using the OD readings provided in the data set (i.e., $Y_{ij} = 1$ if the corresponding individual OD reading exceeded t^*). This entire process leaves us with $B = 500$ DT data sets for analysis. Note that dilution is introduced by assuming the same threshold t^* is used for both master pool tests and individual retests [26].

We consider the population-level model

$$\text{logit}\{\text{pr}(\tilde{Y}_{ij}=1|x_{ij})\}=\beta_0+\beta_1x_{ij}+\beta_2x_{ij}^2, \quad (9)$$

where x_{ij} denotes the standardized age of the i th individual in the j th pool. For each of the 500 DT data sets, we estimate the model in (9) while assuming (incorrectly) that the logistic submodel in (3) describes the effect of dilution. For a baseline comparison, we also estimate (9) using the $N = 1137$ individual testing responses in the data set while assuming $S_e = 0.971$ and $S_p = 1$, which we have calculated using the available biomarker information on each individual. Finally, for each DT data set, we estimate (9) assuming $S_e = 0.971$ is the same for master pools and individual retests; this is the so-called “constant S_e ” model that ignores dilution [20].

Figure 1 shows estimated regression functions, averaged over the $B = 500$ data sets, for the constant S_e , dilution, and individual testing data models. From Figure 1, it is clear that the estimated regression function for the constant S_e model is dramatically influenced by dilution. This estimated function on average is uniformly smaller than the averaged estimated function from individual testing across all ages and master pool sizes c_j , especially so when c_j is larger. On the other hand, even when a logistic dilution submodel is incorrectly assumed, our averaged estimated regression function more closely resembles the individual testing data results. This is especially true for larger master pool sizes where the evidence for

dilution is stronger. For example, the proportion of times (out of 500) our LRT test rejects $H_0 : \lambda = 0$ is 0.15, 0.57, 0.84, and 0.97 for pool sizes $c_j = 2, 4, 6$, and 8, respectively.

6.2. Nebraska infertility prevention project data

Our second data set contains chlamydia test results collected by the Nebraska Public Health Laboratory (NPHL) as part of the infertility prevention project (IPP). The data consist of testing results from different genders (male/female) and specimen types (urine/swab) in 2009 and also several covariates measured for each individual. We use the $N = 14530$ female swab specimens collected in 2009 to illustrate our regression methods because this specimen type is pooled in the neighboring state of Iowa as part of their statewide chlamydia screening efforts [5]. Also, unlike the Irish prisoner example, this data set does not contain continuous biomarker responses.

We create $B = 500$ DT data sets for analysis in a manner similar to that in Section 6.1. Because no biomarker information is available, we treat the individual diagnoses in the data set as the “true” statuses $\tilde{Y}_{ij}$ and then simulate master pool responses $Z_j \sim \text{Bernoulli}\{h(k, c_j, \lambda)I_{jk} + (1 - S_p)I_{j0}\}$ and individual responses $Y_{ij} \sim \text{Bernoulli}\{S_e Y_{ij} + (1 - S_p)\tilde{Y}_{ij}\}$ for master pools that test positively. We use master pools of size $c_j = 5$, because this pool size minimizes the expected number of tests needed to classify each female subject as positive or negative when DT is used [5], and we form pools by random assignment. Individual testing assay accuracy probabilities are assumed to be $S_e = 0.942$ and $S_p = 0.976$. These values are taken directly from the product literature for the Aptima Combo 2 Assay (Hologic, San Diego), which is currently used by the NPHL.

We consider the population-level model

$$\text{logit}\{\text{pr}(\tilde{Y}_{ij}=1|x_{ij})\} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}, \quad (10)$$

where the three covariates x_1 , x_2 , and x_3 denote age (in years), symptoms status (individual shows symptoms/not), and new partner status (new partner in last 6 months/not), respectively. To examine the effect of dilution submodel misspecification with these data, we dilute the master pool responses Z_j using

$$h(k, c_j, \lambda) = \frac{S_e k}{k + \lambda(c_j - k)},$$

the Hung and Swallow (1999) submodel. We then estimate (10) and $h(k, c_j, \lambda)$ under the assumption that this submodel is correct and also under the assumption that the logistic submodel in (3) is correct. Values of $\lambda = 0.05$, $\lambda = 0.10$, and $\lambda = 0.15$ are used with the Hung and Swallow (1999) submodel to generate the master pool testing responses Z_j ; these values are associated with dilution levels that are increasing in λ ; see Table A.3 in the Web-based Supporting information. Finally, we estimate (10) with the $N = 14530$ individual test responses in the data set for comparison.

Table 4 summarizes the results. One notes that regardless of whether the correct dilution submodel is used, our methods provide (averaged) maximum likelihood estimates and standard errors that are nearly identical (under the correct and incorrect dilution submodel) and are also very similar to the same estimates obtained from individual testing. That the standard errors are larger under DT, when compared to individual testing, is not surprising because (a) our dilution regression methods estimate the additional parameter λ and (b) DT uses less than half the number of tests as individual testing (which uses $N = 14530$ tests). Finally, the empirical rejection rates of our LRT for dilution are quite encouraging and are not largely affected by whether the correct submodel is used for estimation.

7. Discussion

We have proposed a new regression methodology that accounts for dilution in group testing without requiring underlying biomarker information. Our approach corrects for estimation bias in traditional regression approaches by using a parametric submodel $h(k, c_j, \lambda)$ that describes dilution. Developing such a submodel is straightforward and can use any cumulative distribution function as shown in Section 2. Furthermore, our LRT in Section 4 allows the investigator to formally test whether dilution exists in the data, an inferential technique not available with previous regression approaches that require biomarker distributions to be known. Our research web site www.chrisbilder.com/grouptesting contains R programs that implement our methods for any choice of submodel under MPT and DT.

Group testing is most commonly applied in high volume clinical specimen settings where the ultimate goal is to classify each subject as positive or negative. The simulations and data applications in this article reveal that our techniques perform well when DT is used to accomplish this. Future work could generalize the methods herein to analyze group testing data from higher-dimensional hierarchical algorithms or from array testing [14, 15]. Both extensions might be formidable, both analytically and computationally, but the estimation methods in Section 3.2 could serve as a starting point towards accomplishing this. Other extensions could include using submodels which are specified nonparametrically; e.g., $h(\cdot, \cdot, \lambda)$ monotone in λ , or using Bayesian methods if prior information on the dilution effect is available. Finally, it should be possible to utilize our dilution submodel approach with other group testing regression methods, such as those which incorporate random effects [36], covariate measurement error [37], and aggregated covariates [38].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to two anonymous referees for providing us with insightful comments on an earlier version of this article. We thank our colleagues at the Nebraska Public Health Laboratory for their collaboration and also Dr. Shane Allwright at Trinity College Dublin for allowing us to use the Irish prisoner data. This research was funded by Grant R01 AI121351 from the National Institutes of Health.

References

1. Dorfman R. The detection of defective members of large populations. *Annals of Mathematical Statistics*. 1943; 14:436–440. DOI: 10.1214/aoms/1177731363
2. Krajden M, Cook D, Mak A, Chu K, Chahil N, Steinberg M, Rekart M, Gilbert M. Pooled nucleic acid testing increases the diagnostic yield of acute HIV infections in a high-risk population compared to 3rd and 4th generation HIV enzyme immunoassays. *Journal of Clinical Virology*. 2014; 61:132–137. <http://dx.doi.org/10.1016/j.jcv.2014.06.024>. [PubMed: 25037533]
3. Stramer S, Wend U, Candotti D, Foster G, Hollinger F, Dodd R, Allain J, Gerlich W. Nucleic acid testing to detect HBV infection in blood donors. *New England Journal of Medicine*. 2011; 364:236–247. DOI: 10.1056/NEJMoa1007644 [PubMed: 21247314]
4. Sarov B, Novack L, Beer N, Safi J, Soliman H, Pliskin J, Litvak E, Yaari A, Shinar E. Feasibility and cost-benefit of implementing pooled screening for HCVAg in small blood bank settings. *Transfusion Medicine*. 2007; 17:479–487. DOI: 10.1111/j.1365-3148.2007.00790.x [PubMed: 17727618]
5. Tebbs J, McMahan C, Bilder C. Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. *Biometrics*. 2013; 69:1064–1073. DOI: 10.1111/biom.12080 [PubMed: 24117173]
6. Busch M, Caglioti S, Robertson E, McAuley J, Tobler L, Kamel H, Linnen J, Shyamala V, Tomasulo P, Kleinman S. Screening the blood supply for West Nile virus RNA by nucleic acid amplification testing. *New England Journal of Medicine*. 2005; 353:460–467. DOI: 10.1056/NEJMoa044029 [PubMed: 16079369]
7. Edouard S, Prudent E, Gautret P, Memish Z, Raoult D. Cost-effective pooling of DNA from nasopharyngeal swab samples for large-scale detection of bacteria by real-time PCR. *Journal of Clinical Microbiology*. 2015; 52:1002–1004. DOI: 10.1128/JCM.03609-14
8. Heffernan A, Aylward L, Toms L, Sly P, Macleod M, Mueller J. Pooled biological specimens for human biomonitoring of environmental chemicals: Opportunities and limitations. *Journal of Exposure Science and Environmental Epidemiology*. 2014; 24:225–232. DOI: 10.1038/jes.2013.76 [PubMed: 24192659]
9. Speybroeck N, Williams C, Lafia K, Devleesschauwer B, Berkvens D. Estimating the prevalence of infections in vector populations using pools of samples. *Medical and Veterinary Entomology*. 2012; 26:361–371. DOI: 10.1111/j.1365-2915.2012.01015.x [PubMed: 22486773]
10. Berger T, Mandell J, Subrahmanya P. Maximally efficient two-stage screening. *Biometrics*. 2000; 56:833–840. DOI: 10.1111/j.0006-341X.2000.00833.x [PubMed: 10985223]
11. Dhand N, Johnson W, Toribio J. A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *Journal of Agricultural, Biological, and Environmental Statistics*. 2010; 15:452–473. DOI: 10.1007/s13253-010-0032-8
12. Remlinger K, Hughes-Oliver J, Young S, Lam R. Statistical design of pools using optimal coverage and minimal collision. *Technometrics*. 2006; 48:133–143. DOI: 10.1198/004017005000000481
13. Fahey J, Ourisson P, Degnan F. Pathogen detection, testing, and control in fresh broccoli sprouts. *Nutrition Journal*. 2006; 5:13. doi: 10.1186/1475-2891-5-13 [PubMed: 16630354]
14. 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. DOI: 10.1111/j.1541-0420.2007.00817.x [PubMed: 17501946]
15. Westreich D, Hudgens M, Fiscus S, Pilcher C. Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests. *Journal of Clinical Microbiology*. 2008; 46:1785–1792. DOI: 10.1128/JCM.00787-07 [PubMed: 18353930]
16. Hepworth G, Watson R. Debiased estimation of proportions in group testing. *Journal of the Royal Statistical Society, Series C*. 2009; 58:105–121. DOI: 10.1111/j.1467-9876.2008.00639.x
17. Liu A, Liu C, Zhang Z, Albert P. Optimality of group testing in the presence of misclassification. *Biometrika*. 2012; 99:245–251. DOI: 10.1093/biomet/asr064 [PubMed: 23049137]
18. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000; 56:1126–1133. DOI: 10.1111/j.0006-341X.2000.01126.x [PubMed: 11129470]

19. Xie M. Regression analysis of group testing samples. *Statistics in Medicine*. 2001; 20:1957–1969. DOI: 10.1002/sim.817 [PubMed: 11427952]
20. Zhang B, Bilder C, Tebbs J. Group testing regression model estimation when case identification is a goal. *Biometrical Journal*. 2013; 55:173–189. DOI: 10.1002/bimj.201200168 [PubMed: 23401252]
21. Wein L, Zenios S. Pooled testing for HIV screening: Capturing the dilution effect. *Operations Research*. 1996; 44:543–569. DOI: 10.1287/opre.44.4.543
22. Zenios S, Wein L. Pooled testing for HIV prevalence estimation: Exploiting the dilution effect. *Statistics in Medicine*. 1998; 17:1447–1467. [PubMed: 9695191]
23. Hwang F. Group testing with a dilution effect. *Biometrika*. 1976; 63:671–673. DOI: 10.1093/biomet/63.3.671
24. Hung M, Swallow W. Robustness of group testing in the estimation of proportions. *Biometrics*. 1999; 55:231–237. DOI: 10.1111/j.0006-341X.1999.00231.x [PubMed: 11318160]
25. McMahan C, Tebbs J, Bilder C. Regression models for group testing data with pool dilution effects. *Biostatistics*. 2013; 14:284–298. DOI: 10.1093/biostatistics/kxs045 [PubMed: 23197382]
26. Wang D, McMahan C, Gallagher C. A general parametric regression framework for group testing data with dilution effects. *Statistics in Medicine*. 2015; 34:3606–3621. DOI: 10.1002/sim.6578 [PubMed: 26173957]
27. Delaigle A, Hall P. Nonparametric methods for group testing data, taking dilution into account. *Biometrika*. 2015; 102:871–887. DOI: 10.1093/biomet/asv049
28. Zhang Z, Liu C, Kim S, Liu A. Prevalence estimation subject to misclassification: The mis-substitution bias and some remedies. *Statistics in Medicine*. 2014; 33:4482–4500. DOI: 10.1002/sim.6268 [PubMed: 25043925]
29. Wang Y. On the number of successes in independent trials. *Statistica Sinica*. 1993; 3:295–312.
30. Louis T. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1982; 44:226–233.
31. Self S, Liang K. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*. 1987; 82:605–610. DOI: 10.1080/01621459.1987.10478472
32. Molenberghs G, Verbeke G. Likelihood ratio, score, and Wald tests in a constrained parameter space. *American Statistician*. 2007; 61:22–27. DOI: 10.1198/000313007X171322
33. Sarker, M. Ph.D. dissertation. University of South Carolina; 2016. Modern Estimation Problems in Group Testing.
34. Bilder C, Tebbs J. Bias, efficiency, and agreement for group-testing regression models. *Journal of Statistical Computation and Simulation*. 2009; 79:67–80. DOI: 10.1080/00949650701608990 [PubMed: 20046956]
35. Allwright S, Bradley F, Long J, Barry J, Thornton L, Parry J. Prevalence of antibodies to hepatitis B, hepatitis C, and HIV and risk factors in Irish prisoners: Results of a national cross sectional survey. *British Medical Journal*. 2000; 321:78–82. DOI: 10.1136/bmj.321.7253.78 [PubMed: 10884256]
36. Chen P, Tebbs J, Bilder C. Group testing regression models with fixed and random effects. *Biometrics*. 2009; 65:1270–1278. DOI: 10.1111/j.1541-0420.2008.01183.x [PubMed: 19210734]
37. Huang X, Tebbs J. On latent-variable model misspecification in structural measurement error models for binary response. *Biometrics*. 2009; 65:710–718. DOI: 10.1111/j.1541-0420.2008.01128.x [PubMed: 18945265]
38. Delaigle A, Zhou W. Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates. *Journal of the American Statistical Association*. 2015; 110:1785–1796. DOI: 10.1080/01621459.2015.1054491

Figure 1.

Irish HBV data. Estimated regression functions for the model in (9), averaged over $B = 500$ implementations of DT for different master pool sizes $c_j = c$. Master pools were created by random assignment. The covariate x is the standardized value of age. In each subfigure, the estimated individual data regression function is shown for comparison.

Logistic dilution submodel. Pool-level sensitivity $h(k, c_j, \lambda) = \text{pr}_\lambda(Z_j = 1 | \sum_{i=1}^{c_j} \tilde{Y}_{ij} = k)$ for different values of the dilution parameter λ when $S_e = 0.99$ and the master pool size $c_j \in \{5, 10\}$. Tables for other dilution submodels are given in the Web-based Supporting information.

Table 1

λ	$k=1$	2	3	4	5	6	7	8	9	10
5	2.6	0.93	0.95	0.97	0.98	0.99				
	3.8	0.83	0.91	0.96	0.98	0.99				
	5.0	0.64	0.83	0.93	0.97	0.99				
10	2.6	0.91	0.93	0.94	0.95	0.96	0.97	0.98	0.99	0.99
	3.8	0.76	0.83	0.87	0.91	0.94	0.96	0.97	0.98	0.99
	5.0	0.52	0.64	0.75	0.83	0.89	0.93	0.96	0.97	0.98

Simulation study for Dorfman testing (DT). Results are based on $B = 500$ simulated group testing data sets from the regression model in (8) under the logistic dilution submodel in (3). “Mean” is the average of the 500 maximum likelihood estimates. “Cov” is the empirical coverage probability of nominal 95% Wald confidence intervals (the margin of error is 0.03). “SD” is the sample standard deviation of the 500 estimates and “SE” is the averaged standard error. The level of dilution is controlled by $\lambda \in \{2.6, 3.8, 5.0\}$; see Table 1. Results for random pooling and homogeneous pooling are shown. The true value is $\beta = (-3, 2, 1)'$.

Table 2

λ	Pool size	Model with constant S_e			Model with logistic dilution submodel		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
2.6	Random	Mean (Cov)	-3.05 (0.91)	1.96 (0.93)	0.98 (0.95)	-3.00 (0.95)	2.00 (0.96)
		SD (SE)	0.09 (0.09)	0.09 (0.09)	0.15 (0.15)	0.10 (0.10)	0.15 (0.16)
	Homogeneous	Mean (Cov)	-3.08 (0.87)	2.03 (0.94)	1.01 (0.95)	-3.01 (0.97)	2.00 (0.95)
		SD (SE)	0.08 (0.09)	0.10 (0.09)	0.15 (0.15)	0.10 (0.10)	0.15 (0.16)
	Random	Mean (Cov)	-3.08 (0.91)	1.96 (0.92)	0.97 (0.95)	-3.00 (0.95)	2.00 (0.94)
		SD (SE)	0.09 (0.09)	0.10 (0.10)	0.15 (0.15)	0.10 (0.10)	0.15 (0.16)
3.8	Homogeneous	Mean (Cov)	-3.10 (0.83)	2.05 (0.91)	1.03 (0.94)	-3.00 (0.95)	2.00 (0.93)
		SD (SE)	0.09 (0.09)	0.10 (0.09)	0.16 (0.15)	0.10 (0.10)	0.16 (0.16)
	Random	Mean (Cov)	-3.15 (0.66)	1.92 (0.86)	0.95 (0.94)	-3.01 (0.96)	2.01 (0.95)
		SD (SE)	0.09 (0.09)	0.09 (0.09)	0.16 (0.15)	0.10 (0.10)	0.17 (0.16)
	Homogeneous	Mean (Cov)	-3.19 (0.46)	2.07 (0.90)	1.03 (0.93)	-3.00 (0.95)	2.00 (0.96)
		SD (SE)	0.09 (0.09)	0.10 (0.10)	0.17 (0.16)	0.11 (0.11)	0.16 (0.16)
5.0	Random	Mean (Cov)	-3.21 (0.40)	1.91 (0.81)	0.95 (0.94)	-3.01 (0.95)	2.01 (0.95)
		SD (SE)	0.10 (0.10)	0.10 (0.10)	0.16 (0.16)	0.11 (0.11)	0.16 (0.17)
	Homogeneous	Mean (Cov)	-3.28 (0.17)	2.15 (0.67)	1.08 (0.89)	-3.02 (0.95)	2.01 (0.95)
		SD (SE)	0.10 (0.09)	0.11 (0.10)	0.18 (0.16)	0.11 (0.11)	0.17 (0.17)
	Random	Mean (Cov)	-3.33 (0.05)	1.85 (0.64)	0.89 (0.91)	-3.01 (0.95)	2.01 (0.96)
		SD (SE)	0.09 (0.10)	0.09 (0.10)	0.17 (0.16)	0.11 (0.11)	0.18 (0.18)
5.0	Homogeneous	Mean (Cov)	-3.44 (0.01)	2.17 (0.60)	1.09 (0.88)	-3.01 (0.94)	2.01 (0.94)
		SD (SE)	0.10 (0.10)	0.12 (0.10)	0.18 (0.16)	0.12 (0.12)	0.17 (0.17)
	Random	Mean (Cov)	-3.49 (0.00)	1.82 (0.58)	0.88 (0.89)	-3.00 (0.95)	2.01 (0.95)
		SD (SE)	0.11 (0.11)	0.10 (0.11)	0.18 (0.18)	0.12 (0.12)	0.20 (0.20)

λ	Pool size	Model with constant S_e			Model with logistic dilution submodel		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
		Mean (Cov)	SD (SE)	Homogeneous	Mean (Cov)	SD (SE)	Homogeneous
		-3.65 (0.00)	0.14 (0.11)	2.36 (0.13)	1.18 (0.76)	-3.02 (0.96)	2.01 (0.94)
				0.14 (0.12)	0.22 (0.18)	0.13 (0.13)	0.12 (0.12)
							1.00 (0.95)
							0.18 (0.18)

Table 3

LRT for dilution under the logistic submodel in (3). Estimated size and power of the LRT based on $B = 500$ group testing data sets from the regression model in (8) with $\alpha = 0.05$. The margin of error of the estimated size when $\lambda = 0$, assuming a 99% confidence level, is 0.03. The amount of dilution increases as λ increases. Results under dilution submodel misspecification are given in the Web-based Supporting information.

c_j	Master pool testing (MPT)					Dorfman testing (DT)				
	$\lambda = 0$	1	2	3	4	$\lambda = 0$	1	2	3	4
Random pooling										
5	0.05	0.08	0.06	0.13	0.22	0.06	0.08	0.16	0.48	0.93
10	0.05	0.09	0.10	0.20	0.27	0.05	0.08	0.37	0.90	1.00
Homogeneous pooling										
5	0.04	0.06	0.08	0.14	0.16	0.05	0.08	0.19	0.49	0.94
10	0.06	0.09	0.20	0.31	0.38	0.03	0.11	0.37	0.85	1.00

Nebraska infertility prevention project data. Mean (averaged standard error) of $B = 500$ maximum likelihood estimates under model (10). Master pools were created by random assignment and DT was used. Dilution is introduced by assuming the Hung and Swallow (1999) submodel, HS (1999), is correct. The proportion of times (out of 500) the LRT test in Section 4 rejects $H_0 : \lambda = 0$ is also shown.

Table 4

	Intercept	Age	Symptoms	New partner	Reject H_0
Individual testing	-0.71 (0.16)	-0.09 (0.01)	0.64 (0.07)	-0.07 (0.09)	--
$\lambda = 0.05$	-0.72 (0.21)	-0.09 (0.01)	0.65 (0.09)	-0.08 (0.11)	0.74
HS (1999)	-0.72 (0.22)	-0.09 (0.01)	0.65 (0.09)	-0.08 (0.11)	0.97
$\lambda = 0.15$	-0.71 (0.24)	-0.09 (0.01)	0.65 (0.10)	-0.08 (0.12)	1.00
$\lambda = 0.05$	-0.68 (0.21)	-0.09 (0.01)	0.65 (0.09)	-0.08 (0.11)	0.72
Logistic	-0.69 (0.23)	-0.09 (0.01)	0.65 (0.09)	-0.09 (0.11)	0.97
$\lambda = 0.15$	-0.67 (0.24)	-0.09 (0.01)	0.65 (0.10)	-0.08 (0.12)	1.00