Published in final edited form as:

J Stat Comput Simul. 2009 January 1; 79(1): 67–80. doi:10.1080/00949650701608990.

Bias, efficiency, and agreement for group-testing regression models

Christopher R. Bilder^{*,†} and Joshua M. Tebbs[‡]

[†]Department of Statistics, University of Nebraska, Lincoln, NE 68583, U.S.A.

[‡]Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A.

Abstract

Group testing involves pooling individual items together and testing them simultaneously for a rare binary trait. Whether the goal is to estimate the prevalence of the trait or to identify those individuals that possess it, group testing can provide substantial benefits when compared to testing subjects individually. Recently, group-testing regression models have been proposed as a way to incorporate covariates when estimating trait prevalence. In this paper, we examine these models by comparing fits obtained from individual and group testing samples. Relative bias and efficiency measures are used to assess the accuracy and precision of the resulting estimates using different grouping strategies. We also investigate the agreement of individual and group-testing regression estimates for various grouping strategies and the effects of group size selection. Depending on how groups are formed, our results show that group-testing regression models can perform very well when compared to the analogous models based on individual observations. However, different grouping strategies can provide very different results in finite samples.

Keywords

Binary data; Diagnostic test; Generalised linear model; Pooling; Prevalence; Unobserved data

1. Introduction

In many studies involving low-prevalence traits, group testing (also known as pooled testing) is used to increase efficiency and to reduce testing costs. In such contexts, individual items (e.g., blood/urine specimens, drug compounds, gene segments, insect vectors, etc.) are pooled together and tested simultaneously rather than individually. Group testing has a rich history dating back to the works of Watson [1] and Dorfman [2]. Today, group testing is exploited in a variety of areas including epidemiology, genetics, blood-bank screening, drug discovery, biology, and plant pathology [3-9]. Depending on the application, group testing may be used simply to estimate the proportion of positives or to classify each individual item as positive or negative. Our focus is on the estimation problem within a heterogeneous population.

Using pooled samples in epidemiological investigations is now a ubiquitous approach when estimating the prevalence of sexually-transmitted diseases like HIV, hepatitis, syphilis, chlamydia, and gonorrhea [10-12], and the savings conferred from pooling in these applications are well-documented. Pooling is a safe alternative to testing each subject individually. Traditional ELISA tests used in HIV screening are reported to have negligible loss in sensitivity and specificity for pool sizes up to 15, and rapid antibody assays are shown to have no loss in

*Corresponding author. chris@chrisbilder.com.

accuracy for pools up to size 20 [13-14]. The use of nucleic acid amplification testing (NAT) technology, such as the polymerase chain reaction (PCR) assay, is now widespread in Europe and North America. These assays detect viral nucleic acids directly rather than an individual subject's antibodies. NAT technology safely allows for large pool sizes to be used and also dramatically reduces the diagnostic window period [15-16].

To model the effects of covariates with group-testing data, some authors have proposed using a regression approach. Farrington [17] was the first to suggest modeling covariate-specific probabilities for individuals in groups with identical covariate values, using a generalised linear model (GLM) for binomial data with a complementary log-log link function. Vansteelandt et al. [18] and Xie [19] generalise this work by allowing for arbitrary link functions and different covariate values in the pools. Xie [19] also discusses the possibility of using regression models outside a GLM setting. However, within a GLM context, both approaches in [18] and [19] produce maximum likelihood estimates of the model parameters. Vansteelandt et al. [18] use a method by which maximum likelihood estimates are computed directly using the group responses. Xie [19] treats the individual responses as unobserved and uses an expectation-maximisation (EM) algorithm to find the maximum likelihood estimates. Tu et al. [20] suggests a Bayesian approach for modeling, but like Farrington [17], make the restrictive assumption that covariate values are identical within groups.

Modeling group-testing data has been a notable advance, because it allows one to take advantage of risk factors through the covariates while acknowledging population heterogeneity. This notwithstanding, implementing this methodology now presents new challenges for public-health officials, lab technicians, and data analysts. For example, perhaps the most important initial design consideration is to decide exactly how the pools are formed. Vansteelandt et al. [18] show that pools with homogenous individuals (in terms of their covariate values) will minimise the variance of model-based estimators. However, while constructing homogenous pools confers this advantage, we have found that this pooling strategy may not be practical. For example, in infectious-disease screening, individual samples of blood or sera have a finite shelf life [21] so one can not simply wait for other homogenous individuals to enter the study. In drug discovery, Remlinger et al. [7] argue that constructing homogenous pools may be suboptimal if the goal is to identify individual compounds which are biologically active. In entomology, Katholi and Unnasch [22] explain that researchers often have no choice in determining how to pool insect vectors. As these examples illustrate, homogenous pooling is not always possible nor is it always an appropriate design. In this light, one might naturally wonder if non-homogenous pooling strategies would ultimately lead to different conclusions about covariate effects.

In this paper, we investigate the effects of non-homogenous pooling strategies on parameter estimation in group-testing regression models. Throughout, we compare these strategies to both homogenous pooling and individual testing to determine what is gained or lost from implementing them. In Section 2, we define our notation, summarise the approaches of Vansteelandt et al. [18] and Xie [19], and discuss different pooling strategies. In Section 3, we discuss a Kenyan HIV example and use it as a basis for an initial investigation. In Section 4, we then provide a detailed simulation study describing the effects of different pooling strategies and different pool sizes. In Section 5, we conclude with a summary and a discussion on future research topics. Throughout this paper, we use the terms “group” and “pool” interchangeably.

2. Background

In the group-testing literature, various protocols have been proposed for data collection. Hammick and Gastwirth [23] propose a strategy wherein two sets of pools are initially formed so that each individual is (indirectly) tested twice. This strategy is shown to be useful when

estimating traits with moderate prevalence levels. Phatarfod and Sudbury [24] discuss pooling experiments which use a matrix-array testing scheme. This design is particularly amenable to testing blood samples in a highly-automated environment that implements NAT screening. Gastwirth and Johnson [25] and Johnson and Gastwirth [26] propose protocols where individuals and/or groups are tested in two stages; this approach (and variants thereof) has been shown to serve as an effective quality-control tactic and to reduce the number of misclassified individuals. Finally, Xie et al. [8] examine data from a drug-discovery experiment where each individual chemical compound is initially put into two groups using a square-array approach. These authors also consider a trinomial response that includes positive, negative, and blocker responses (although not in a regression modeling context).

While each of the aforementioned pooling protocols is meritorious in its own right, in this paper, we specifically focus on the protocol wherein a single set of pools is available and each pool contributes only one response. This is arguably the simplest design to implement, but it is the most pervasive design in the infectious-disease, public-health, and plant-pathology literatures. To cite only a few examples, the American Red Cross uses this protocol to screen blood donations for infectious disease surveillance programmes [27,28]. Public-health observational studies use this protocol to monitor gonorrhea and chlamydia prevalence [29, 30]. Out of necessity, entomologists and plant pathologists use this protocol when testing pools of insect vectors for arthropod-borne infections [22,31,32]. We remark on other group-testing data collection methods in Section 5.

2.1. Notation and assumptions

We define $Y_{ik} = 1$ if the i th item in the k th group (i.e., pool) is positive; $Y_{ik} = 0$ otherwise, for $i = 1, 2, \dots, I_k$ and $k = 1, 2, \dots, K$. In a group-testing context, measuring a “trait” produces a binary response, such as the presence/absence of HIV infection. An “item” refers to experimental material that can be tested for each subject, such as a blood or a urine specimen. We assume that all individual responses, Y_{ik} , are independent Bernoulli random variables with mean $p_{ik} = P(Y_{ik} = 1)$ and let Z_k denote the binary group response for the k th group; i.e., $Z_k = 1$ if the group tests positive for the trait and $Z_k = 0$ if the group tests negative. We also assume that the Z_k are independent Bernoulli random variables with mean $\theta_k = P(Z_k = 1)$. Provided that groups are not misclassified, one will note that $Z_k = 1$ if and only if $\sum_{i=1}^{I_k} Y_{ik} > 0$ (i.e., at least one item in the group is positive) and $Z_k = 0$ if and only if $\sum_{i=1}^{I_k} Y_{ik} = 0$. Thus, in a profound sense, the individual statuses $Y_{1k}, Y_{2k}, \dots, Y_{I_k k}$ are indirectly “observed” when $Z_k = 0$; they are unobserved otherwise.

We define the sensitivity and the specificity of a group test by $\eta = P(Z_k = 1 | \tilde{Z}_k = 1)$ and $\delta = P(Z_k = 0 | \tilde{Z}_k = 0)$, respectively, where $\tilde{Z}_k$ denotes the true positive/negative status of group k . Using a conditional probability argument, one can show that $\theta_k = \eta + (1 - \delta - \eta)P(Z_k = 0)$. Both Vansteelandt et al. [18] and Xie [19] assume that η and δ are known, are common for all groups, and do not depend on the group size I_k . These are often reasonable assumptions provided that I_k is not too large. Vansteelandt et al. [18, pp 1127] reports that $\eta = 1$ and $\delta = 0.9997$ for an assay test used in a Kenyan study that we discuss in Section 3. To simplify the comparisons we provide in Sections 3 and 4, we henceforth take $\eta = 1$ and $\delta = 1$ so that

$$\theta_k = 1 - \prod_{i=1}^{I_k} (1 - p_{ik}). \quad (1)$$

Some may view this assumption as confining, but it is important to remember that our focus here is to explore the usefulness of various pooling strategies and to investigate the comparative merits of individual and group testing models. Both individual testing and group testing designs are subject to misclassification, and empirical evidence suggests that these misclassification rates are similar for both designs [13-16,18].

2.2. Regression models

Let $x_{ik} = (1, x_{ik1}, x_{ik2}, \dots, x_{ikr})'$ denote an $(r + 1) \times 1$ vector of covariate values for the i th item in the k th group, and let $\beta = (\beta_0, \beta_1, \dots, \beta_r)'$ denote the corresponding vector of model parameters. A statistical model for p_{ik} can be expressed as $f(p_{ik}) = \beta' x_{ik}$. In this manuscript, we restrict attention to a generalised linear models framework and the commonly-used logit function $f(w) = \log[w/(1 - w)]$, for $0 < w < 1$, so that the model of interest becomes

$$\log \left(\frac{p_{ik}}{1 - p_{ik}} \right) = \beta' x_{ik}. \quad (2)$$

As noted earlier, Vansteelandt et al. [18] and Xie [19] both consider models of the form in (2), but present different ways to find the maximum likelihood estimator (MLE) of β , which we denote by $\hat{\beta}_G$. Vansteelandt et al. [18] expresses the likelihood function in terms of the observed z_k responses only; i.e.,

$$\begin{aligned} L \equiv L(\beta) &= \prod_{k=1}^K \theta_k^{z_k} (1 - \theta_k)^{1-z_k} \\ &= \prod_{k=1}^K \left[1 - \prod_{i=1}^{I_k} (1 - p_{ik}) \right]^{z_k} \left[\prod_{i=1}^{I_k} (1 - p_{ik}) \right]^{1-z_k}, \end{aligned}$$

for the model in (2), and maximises L directly to find the MLE $\hat{\beta}_G$. The covariance matrix for $\hat{\beta}_G$ can be written in closed form using standard calculations from maximum likelihood estimation theory; see Appendix A in [18]. Xie [19] takes an alternative approach by expressing the likelihood function in terms of the unobserved individual responses, which allows one to consider more complicated group-testing data collection methods. Conceptualising the problem in this way, maximum likelihood estimates of the parameters can be obtained using the EM algorithm. However, because we are focused on the simplest data collection protocol as noted earlier, we will use the estimation approach outlined in [18].

2.3. Pooling strategies

Before Farrington's [17] seminal paper, all of the statistical work in the estimation problem treated individuals as homogenous. In a homogenous population, the basic assumption is that individual items are randomly assigned to groups. This ensures that the group responses are independent and identically distributed random variables and allows the MLE of the common prevalence p to be calculated in closed form; see, e.g., [33]. However, in a heterogeneous population where an individual's covariate information is available and could be exploited, there are different possible pooling strategies at one's disposal. We focus on three such strategies, which we term "alike," "random," and "different." Vansteelandt et al. [18] argues that $\hat{\beta}_G$ is a consistent estimator of β regardless of which pooling strategy is used, so long as pools are randomly composed conditional on the covariates.

2.3.1. The “alike” pooling strategy—With the alike (AL) strategy, individuals with “similar” covariates are grouped together, making the pools as homogenous as possible. As noted earlier, Vansteelandt et al. [18] demonstrate that this pooling strategy provides the smallest variances for the components of β_G . When there is only one covariate (i.e., $r = 1$), one can simply order the covariate values from low to high and form groups consecutively based on this ordering. When there is more than one covariate (i.e., $r > 1$), Vansteelandt et al. [18] suggest an ad-hoc approach whereby one first sorts by the “most important” covariate. Then, one subsequently sorts the values of the second most important covariate within alike levels of the first. This process continues until all covariates have been sorted.

The AL strategy is intuitively appealing and provides the best model-based estimates; however, using this pooling strategy requires one to have all individual items up front and available for testing at the same time. This was the case, for example, in Thorburn et al. [34] who estimated hepatitis C virus (HCV) prevalence among health care workers in western Scotland using pooled blood samples. (As a side note, the AL pooling strategy was not used by the authors.) In other investigations, it may not be prudent to assume that all individuals are available for testing at the same time. It may be more likely that individuals are recruited over time, enter the study in a staggered-entry fashion, and are assigned to pools as they become available. This was the case in the Kenyan HIV study examined in [18]. Therefore, while the AL strategy may be the best for parameter estimation, its implementation may not be possible.

2.3.2. The “random” pooling strategy—The “random” (RA) pooling strategy presumes that individual items are randomly assigned to pools irrespective of their covariate values. Of course, to guarantee a true random allocation of individuals to pools, it would be best to have all individual specimens up front; in fact, this was the pooling strategy used by Thorburn et al. [34]. However, assigning individuals to pools in a chronological manner may very well emulate a random assignment, provided that there is no confounding with time [35]. From a practical standpoint, forming the groups in this way may be necessary due to the shelf-life restrictions or perhaps simply for ease of implementation.

2.3.3. The “different” pooling strategy—The “different” (DI) pooling strategy assigns individual items to groups where the covariate values are as disparate as possible. Insofar as model estimation, one might posit that this strategy serves as a “worst-case” scenario. However, there are group-testing applications where forming pools in this way is preferred. As noted earlier, in screening chemical libraries during early phases of the drug-discovery process, Remlinger et al. [7] shows that compound pools with “minimal collision” (i.e., pools containing compounds whose chemical descriptors are as diverse as possible) are best for identifying distinct active compounds. Thus, examining the DI pooling strategy helps us to examine a possible worst-case scenario for parameter estimation and a “best-case” scenario for identification purposes.

3. Kenyan HIV example

Verstraeten et al. [36] and Vansteelandt et al. [18] discuss a public-health study investigating HIV prevalence in rural Kenya. To summarise the salient aspects from [36], blood samples were collected on 740 pregnant women at three rural locations and several potential risk covariates were measured on each individual, including age, marital status, parity, and education level. Blood samples were initially tested for HIV in 74 pools, each of size 10, using ELISA tests. The goals of the analysis in [18] were to determine if the probability of HIV seropositivity for individual subjects was influenced by the potential risk factors, to evaluate the cost-effectiveness of pooling, and to identify an ideal pool size. In order to motivate and explain our main simulation results described in Section 4, we provide a simulated example based on this study.

Vansteelandt et al. [18] consider the simple logistic regression model using age as a covariate; i.e.,

$$\log \left(\frac{p_{ik}}{1 - p_{ik}} \right) = \beta_0 + \beta_1 x_{ik}. \quad (3)$$

Here, x_{ik} denotes the age of the i th individual in the k th group. One of the goals of the Kenyan study was to evaluate the cost-effectiveness of pooling, so investigators in the field collected both the individual (y_{ik}) and the group (z_k) responses. Using both types of data is ideal for our purposes, because we can examine the quality of the regression fits for both individual and group-testing designs. Denote by $\hat{\beta}_G = (\hat{\beta}_{G0}, \hat{\beta}_{G1})'$ and $\hat{\beta}_I = (\hat{\beta}_{I0}, \hat{\beta}_{I1})'$ the maximum likelihood estimators under the model in (3) for grouped and individual data, respectively. Using the observed individual data, Vansteelandt et al. [18] find $\hat{\beta}_{I0} = -1.97$ and $\hat{\beta}_{I1} = -0.024$ from fitting the model in (3). We should remark that there is not a statistically significant relationship between the age term and HIV seropositivity based on the individual observed data. However, this should not detract from our investigation which follows.

To assess the potential benefit from pooling in a regression context, our specific interest lies in comparing the slope estimators $\hat{\beta}_{G1}$ and $\hat{\beta}_{I1}$. The precision of these estimators can be measured by their asymptotic variances. To help simplify the exposition, we henceforth consider only the equal-group size case ($I_k = I$) so that the total number of individuals under consideration is $I \times K$. For $\hat{\beta}_{G1}$, the asymptotic variance is equal to

$$\text{av}(\hat{\beta}_{G1}) = \left\{ \left[\sum_{k=1}^K \gamma_k \left(\sum_{i=1}^I p_{ik} x_{ik} \right)^2 \right] - \frac{\left[\sum_{k=1}^K \gamma_k \left(\sum_{i=1}^I p_{ik} \right) \left(\sum_{i=1}^I p_{ik} x_{ik} \right) \right]^2}{\sum_{k=1}^K \gamma_k \left(\sum_{i=1}^I p_{ik} \right)^2} \right\}^{-1}, \quad (4)$$

where $\gamma_k = \theta_k(1-\theta_k)^{-1}$ and θ_k is defined as in (1). For individual testing, the asymptotic variance of $\hat{\beta}_{I1}$ can be expressed as

$$\text{av}(\hat{\beta}_{I1}) = \left\{ \left[\sum_{k=1}^K \sum_{i=1}^I p_{ik} (1 - p_{ik}) x_{ik}^2 \right] - \frac{\left[\sum_{k=1}^K \sum_{i=1}^I p_{ik} (1 - p_{ik}) x_{ik} \right]^2}{\sum_{k=1}^K \sum_{i=1}^I p_{ik} (1 - p_{ik})} \right\}^{-1}. \quad (5)$$

Equation (5) may look deceiving since $I = 1$ with individual testing; however, the expression is correct when expressed using notation set up for group testing (with individual testing, K would denote the total number of individuals).

We simulate individual and group observations from (3) through the process outlined below. This process allows one to make direct comparisons between individual and group testing models on the same data set.

1. Treat the individual-data estimates $\hat{\beta}_{I0} = -1.97$ and $\hat{\beta}_{I1} = -0.024$ as the true values of β_0 and β_1 , respectively.
2. Simulate x_{ik} from a $\text{Gamma}(\hat{\alpha}, \hat{\beta})$ distribution, where $\hat{\alpha} = 20.95$ and $\hat{\beta} = 1.16$ are the maximum likelihood estimates under the gamma model assumption. This distribution provides a very good fit for the individual age data.

3. Simulate the individual responses Y_{ik} according to a Bernoulli distribution with mean $p_{ik} = \exp(-1.97 - 0.024x_{ik})[1 + \exp(-1.97 - 0.024x_{ik})]^{-1}$.
4. Assign the simulated Y_{ik} values to groups according to a pooling strategy (AL, RA, or DI). For each pool, compute the corresponding group observation

$$Z_k = I \left(\sum_{i=1}^I Y_{ik} > 0 \right).$$

Since there is only one covariate of interest (age), in Step 4 we assign individuals to groups as follows. First, we order the simulated covariate values from low to high. For the AL strategy, the smallest I covariate values are assigned to the first pool, the next smallest I values are assigned to the second pool, and so on, until all individuals have been assigned. For the DI strategy, pool assignments are made by selecting the first available individual in each of I successively-formed clusters. For example, with $K = 100$ and $I = 7$, the first AL group is $x^{(1)}, x^{(2)}, \dots, x^{(7)}$, where $x^{(a)}$ denotes the a th ordered covariate value; the first DI group is $x^{(1)}, x^{(101)}, \dots, x^{(601)}$. RA groups are formed by random assignment.

Table 1 and Figure 1 show the results of fitting the logistic regression model in (3) to just one simulated data set consisting of 700 individuals with $K = 100$ and $I = 7$ using the outlined algorithm and the pooling strategies we have described. These choices of K and I are motivated by the example given in [18, pp 1129]. We used R to fit all models; the `glm()` function was used to fit the individual responses, and the `optim()` function was used to fit the group responses. Table 1 contains the estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, and the estimated asymptotic variance $\widehat{av}(\hat{\beta}_1)$ for each of the three pooling strategies and for individual testing; see Equations (4) and (5). In addition, we report the estimated asymptotic relative efficiency

$$\widehat{re}(\hat{\beta}_{I1} \text{ to } \hat{\beta}_{G1}) = \frac{\widehat{av}(\hat{\beta}_{I1})}{\widehat{av}(\hat{\beta}_{G1})}.$$

This measure is pooling-strategy dependent; values less than unity suggest that individual testing is more efficient. One will note that there is general agreement in the negative relationship between HIV seropositivity and age for each pooling strategy (see Figure 1). That the relative efficiencies for AL, RA, and DI are all less than unity is not surprising because there are 700 individual responses and only 100 group responses. In particular, for this data set, the AL pooling design is only 70 percent as efficient as individual testing, but requires 600 fewer tests. Despite a significant reduction in the number of tests, the RA and DI assignments are much less efficient than individual testing.

In order to provide a more thorough comparison, and to help motivate our simulation study in Section 4, we have used the strategy outlined above to simulate 100 data sets. Table 2 and Figure 2 display the results. The box plots in Figure 2 display the distribution of the $\hat{\beta}_1$ values for the 100 model fits. One notes that the AL strategy produces $\hat{\beta}_1$ estimates that are only marginally more variable than those obtained from individual testing, while the RA and DI strategies produce results that are much more disparate. This can also be seen in Table 2 by looking at the sample variances of the 100 estimates of $\hat{\beta}_1$. In order to measure the agreement between the slope estimates on the same data, we have constructed a parallel coordinates plot in Figure 2 and have computed Pearson correlations in Table 2. One would hope that all pooling strategies would produce approximately the same slope estimates, but this is not the case in this setting. While the AL strategy and individual testing estimates mostly coincide, the RA and DI strategies can produce largely dissimilar estimates of $\hat{\beta}_1$ when compared to individual testing, and even to each other. This lack of agreement is especially noticeable when looking

at the parallel coordinates plot; in fact, for one data set, the largest RA estimate actually coincides with the smallest DI estimate!

4. Comparisons and further simulation evidence

In Section 3, we focused explicitly on the HIV example outlined in Verstraeten et al. [36] and Vansteelandt et al. [18]. In this section, we present a more general comparison. To evaluate the merit of group and individual testing designs in a regression context, there are two obvious bases for comparison. One can compare the two designs (a) when fixing the number of individuals at $I \times K$ (as in Section 3) or (b) when fixing the number of tests at K . Deciding which constraint is appropriate depends on the application. For example, when the number of individual subjects $I \times K$ is limited, perhaps due to recruiting or time constraints, it is this quantity which is best considered fixed. On the other hand, when the number of assays is fixed to manage costs, comparisons with a fixed K are appropriate (this is the situation where group testing is most often preferred). We consider both cases in Sections 4.1 and 4.2, respectively. We follow the simulation algorithm outlined in Section 3, except that we take $\beta_0 = -2$, $\beta_1 = 0.6931$, and generate x_{ik} according to a $\mathcal{U}(-7.079, 1.663)$ distribution. These choices ensure that $0.001 < p_{ik} < 0.30$ and produce an average value of $p_{ik} = 0.02$. Simulation results under different settings and additional measures of agreement (e.g., canonical correlations) are available online at www.chrisbilder.com/grouptesting.

4.1. Fixed number of individuals comparisons

When comparing the two designs with a fixed number of individuals, rarely will group testing outperform individual testing. This is true because individual testing uses $I \times K$ observations while group testing uses only K . In the light of this, we find it necessary to investigate how drastic this loss might be for modeling purposes. As before, central interest will be on the slope estimator $\hat{\beta}_1$. The estimator is asymptotically unbiased; our interest is to assess the amount of bias present with finite samples. For our simulation study, we define the percent bias in $\hat{\beta}_1$ to be

$$\text{pbias}(\hat{\beta}_1) = \beta_1^{-1} \left(B^{-1} \sum_{b=1}^B \hat{\beta}_{1,b} - \beta_1 \right) \times 100,$$

where B denotes the number of simulations and $\hat{\beta}_{1,b}$ denotes the estimate for β_1 using the b th simulated data set. By defining bias in this way, we take into account the relative size of β_1 . Similarly, we define the average estimated asymptotic relative efficiency of $\hat{\beta}_1$ to be

$$\widehat{\text{are}}(\hat{\beta}_{I1} \text{ to } \hat{\beta}_{G1}) = B^{-1} \sum_{b=1}^B \frac{\widehat{\text{av}}(\hat{\beta}_{I1,b})}{\widehat{\text{av}}(\hat{\beta}_{G1,b})},$$

where $\hat{\beta}_{I1,b}$ and $\hat{\beta}_{G1,b}$ denote the b th estimate for β_1 using individual and group testing, respectively. Finally, we define $\widehat{\text{corr}}(\hat{\beta}_{I1}, \hat{\beta}_{G1})$ to be the Pearson correlation for the B simulated slope estimates $(\hat{\beta}_{I1,1}, \hat{\beta}_{G1,1}), (\hat{\beta}_{I1,2}, \hat{\beta}_{G1,2}), \dots, (\hat{\beta}_{I1,B}, \hat{\beta}_{G1,B})$. This measure allows us to assess the empirical agreement among the slope estimates for both designs. Ideally, one would hope for the agreement to be very strong because the estimates are found using the same data sets. However, it becomes clear that this is not necessarily the case.

Table 3 provides the values of $\text{pbias}(\hat{\beta}_1)$ for the three pooling strategies and for individual testing, using different values of I and $I \times K$ with $B = 500$ simulated data sets. As expected, the percent bias decreases as $I \times K$ increases. The pooling strategies generally have a larger percent bias than individual testing with few exceptions occurring for small group sizes (e.g., $I = 2$ and $I = 5$). Overall, bias increases as I does; the only exception to this increasing-bias pattern occurs when $I = 20$ and $I \times K = 200$ for the AL pooling strategy. The reason for this anomalous observation (9.1%) stems from a problem with the convergence of the model-fitting procedure (we exclude all data sets where non-convergence occurs). For these settings, non-convergence rates for the AL, RA, and DI strategies were 44.8, 6.8, and 7.8 percent, respectively. These high rates (particularly for AL) are due largely to “complete separation,” because there are only 10 group observations. Succinctly put, complete separation means that the covariate perfectly discriminates between the binary group responses; this, in turn, creates problems when calculating $\hat{\beta}_G$. For more information on complete separation, see [37]. From our investigation, non-convergence rarely occurs for other settings of I and $I \times K$. However, when the number of groups K is small, investigators should be aware that there may be problems with the model-fitting procedure, especially when similar individuals are placed into common pools.

Table 4 provides values of $\widehat{\text{are}}(\hat{\beta}_I \text{ to } \hat{\beta}_{GI})$ and $\widehat{\text{corr}}(\hat{\beta}_I, \hat{\beta}_{GI})$ for the same settings as in Table 3. Among the pooling strategies, AL is the most efficient as expected. It is interesting to note the sharp decline in group-testing efficiency as I increases, but increasing $I \times K$ from 200 to 1000 does not change the values of $\widehat{\text{are}}(\hat{\beta}_I \text{ to } \hat{\beta}_{GI})$ too drastically. The AL strategy produces slope estimates that are in very good agreement with those obtained from individual testing when I is small. For larger I , this agreement lessens. The RA and DI strategies produce positive correlations but they can be much weaker. As these results suggest, one should not necessarily expect individual and group testing regression models to lead to similar values of the slope estimate $\hat{\beta}_1$, unless perhaps when small group sizes are used.

4.2. Fixed number of tests comparisons

In most applications where group testing is used, costs associated with obtaining individual items are small when compared to the cost of testing. In HIV applications, for example, testing costs may absorb the cost of the assay, the laboratory time for analysis, storage costs, and personnel requirements. In such settings when one is trying to control costs, comparing the two designs on the basis of a fixed number of tests is preferred. Table 5 contains values of $\text{pbias}(\hat{\beta}_1)$ and

$$\widehat{\text{are}}(\hat{\beta}_I \text{ to } \hat{\beta}_{GI}) = \frac{\sum_{b=1}^B \widehat{\text{av}}(\hat{\beta}_{I,b})}{\sum_{b=1}^B \widehat{\text{av}}(\hat{\beta}_{GI,b})}$$

for different group sizes I , numbers of tests K , and $B = 500$. This definition of $\widehat{\text{are}}(\hat{\beta}_I \text{ to } \hat{\beta}_{GI})$ differs slightly from that in Section 4.1 because now we are treating K as fixed, not $I \times K$ (same simulated data set comparisons are not possible in this case). The percentage bias for all three pooling strategies decreases as K increases and maintains a notable advantage over individual testing in nearly all instances. From these results (and other simulation results not shown), we find that the AL and RA strategies are preferred to individual testing when $I < 40$. The DI strategy can provide higher bias than individual testing when I becomes too large (e.g., $I = 10$ when $K = 500$). One also notes from Table 5 that there are gains in efficiency with the group testing models. These gains are quite large for the AL assignment; for RA, the gains are somewhat less. On the other hand, the DI strategy is often less efficient than individual testing.

5. Discussion

This paper investigates the practical issues involved with fitting binary regression models to pooled data. As the AL pooling method can be impractical to implement, we show what is gained or lost from using other strategies. When the number of individuals is fixed, group testing with AL pooling can provide comparable results to individual testing while substantially reducing costs. When fixing the number of tests instead, both AL and RA methods can provide large benefits in terms of bias reduction and increased efficiency. This is reassuring because RA pooling is more practical for use in the field. While the DI strategy appears to perform the worst among the three pooling strategies, it still may be helpful in the classification problem, similarly to that observed by Remlinger et al. [7]. Further research could explore how different pooling strategies are used in a dual-group protocol (possibly using different strategies in different stages) or in a squared-array setting.

In this paper, we have restricted attention to single-covariate regression models in order to focus on the comparisons among different pooling strategies. We would expect to see similar performances for the pooling strategies with multiple-covariate models. However, there are a variety of ways that one could construct AL and DI groups with multiple covariates, and we feel that this topic is worthy of future research. One relatively straightforward approach would be to first use an *I*-means clustering algorithm to form clusters of homogenous individuals (in terms of their covariate structure). Then, individuals could be randomly selected from within or across the *I* clusters to form AL or DI pools, respectively. Also, an anonymous referee has suggested that it may be beneficial to pool individuals sequentially using a measure of covariate similarity. This approach could be fruitful because one could potentially realize the statistical benefits from AL pooling without having to wait until the end of the study.

In homogenous populations, group size determination has received a large amount of attention in the literature. This is an important area since the benefits from group testing depend largely on the size of the pools. In heterogenous populations, the problem becomes much more complex. Vansteelandt et al. [18] provides one way to find optimal group sizes, but their method requires one to have access to all individual items up front; as noted earlier, this may not be possible. Determining optimal group sizes for the different pooling strategies, possibly while adjusting for time confounding as individuals enter the study, would be desirable for technicians in the field.

Acknowledgments

We are indebted to Dr. Stijn Vansteelandt and his colleagues for providing us with the Kenyan data. We also greatly appreciate the comments and recommendations from an anonymous referee.

References

1. Watson M. Factors affecting the amount of infection obtained by aphis transmission of the virus Hy. III. Philosophical Transactions of the Royal Society of London, Series B 1936;226:457–489.
2. Dorfman R. The detection of defective members of large populations. Annals of Mathematical Statistics 1943;14:436–440.
3. Cardoso M, Koerner K, Kubanek B. Mini-pool screening by nucleic acid testing for hepatitis B virus, hepatitis C virus, and HIV: Preliminary results. Transfusion 1998;38:905–907. [PubMed: 9767739]
4. Drosten C, Weber M, Seifried E, Roth W. Evaluation of a new PCR assay with competitive interval control sequence for blood donor screening. Transfusion 2000;40:718–724. [PubMed: 10864995]
5. Pfeiffer R, Rutter J, Gail M, Stuewing J, Gastwirth J. Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. Genetic Epidemiology 22:94–102. [PubMed: 11754476]

6. Tebbs J, Bilder C. Confidence interval procedures for the probability of disease transmission in multiple-vector transfer designs. *Journal of Agricultural, Biological, and Environmental Statistics* 2004;9:75–90.
7. Remlinger K, Hughes-Oliver J, Young S, Lam R. Statistical design of pools using optimal coverage and minimal collision. *Technometrics* 2006;48:133–143.
8. Xie M, Tatsuoka K, Sacks J, Young S. Group testing with blockers and synergism. *Journal of the American Statistical Association* 2000;96:92–102.
9. Thierry-Mieg N. Pooling in systems biology becomes smart. *Nature Methods* 2006;3:161–162. [PubMed: 16489330]
10. 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. [PubMed: 10884256]
11. Aseffa A, Ishak A, Stevens R, Fergusson E, Giles M, Yohannes G, Kidan K. Prevalence of HIV, syphilis, and genital chlamydial infection among women in North-West Ethiopia. *Epidemiology and Infection* 1998;120:171–177. [PubMed: 9593487]
12. Lindan C, Mathur M, Kumta S, Jerajani H, Gogate A, Schachter J, Moncada J. Utility of pooled urine specimens for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR. *Journal of Clinical Microbiology* 2005;43:1674–1677. [PubMed: 15814983]
13. Tu X, Litvak E, Pagano M. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika* 1995;82:287–297.
14. Soroka S, Granade T, Phillips S, Parekh B. The use of simple, rapid tests to detect antibodies to human immunodeficiency virus types 1 and 2 in pooled serum specimens. *Journal of Clinical Virology* 2003;27:90–96. [PubMed: 12727534]
15. Beld M, Habibuw M, Rebers S, Boom R, Reesink H. Evaluation of automated RNA-extraction technology and a qualitative HCV assay for sensitivity and detection of HCV RNA in pool-screening systems. *Transfusion* 2000;40:575–579. [PubMed: 10827262]
16. LeCorfec W, LePont F, Tuckwell H, Rouzioux C, Costagliola D. Direct HIV testing in blood donations: Variation of the yield with detection threshold and pool size. *Transfusion* 1999;39:1141–1144. [PubMed: 10532610]
17. Farrington C. Estimating prevalence by group testing using generalized linear models. *Statistics in Medicine* 1992;11:1591–1597. [PubMed: 1359621]
18. Vansteelandt S, Goetghebuer E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* 2000;56:1126–1133. [PubMed: 11129470]
19. Xie M. Regression analysis of group testing samples. *Statistics in Medicine* 2001;20:1957–1969. [PubMed: 11427952]
20. Tu X, Kowalski J, Jia G. Bayesian analysis for prevalence with covariates using simulation-based techniques: Applications to HIV screening. *Statistics in Medicine* 1999;18:3059–3073. [PubMed: 10544306]
21. Heddle N, Barty R, Sigouin C, Boye D, Nelson E, Blajchman M, Kelton J. In vitro evaluation of prestorage pooled leukoreduced whole blood-derived platelets stored for up to 7 days. *Transfusion* 2005;45:904–910. [PubMed: 15934988]
22. Katholi C, Unnasch T. Important experimental parameters for determining infection rates in arthropod vectors using pool screening approaches. *American Journal of Tropical Medicine and Hygiene* 2006;74:779–785. [PubMed: 16687680]
23. Hammick P, Gastwirth J. Group testing for sensitive characteristics: Extension to higher prevalence levels. *International Statistical Review* 1994;63:319–331.
24. Phatarfod R, Sudbury A. The use of a square array scheme in blood testing. *Statistics in Medicine* 1994;13:2337–2343. [PubMed: 7855467]
25. Gastwirth J, Johnson W. Screening with cost effective quality control: Potential applications to HIV and drug testing. *Journal of the American Statistical Association* 1994;89:972–981.
26. Johnson W, Gastwirth J. Dual group screening. *Journal of Statistical Planning and Inference* 2000;83:449–473.

27. Dodd R, Notari E, Stramer S. Current prevalence and incidence of infectious disease markers and estimated window-period risk in the American Red Cross donor population. *Transfusion* 2002;42:975–979. [PubMed: 12385406]
28. Stramer S. Pooled hepatitis B virus DNA testing by nucleic acid amplification: Implementation or not. *Transfusion* 2005;45:1242–1246. [PubMed: 16078908]
29. Lindan C, Mathur M, Kumta S, Jerajani H, Gogate A, Schachter J, Moncada J. Utility of pooled urine specimens for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR. *Journal of Clinical Microbiology* 2005;43:1674–1677. [PubMed: 15814983]
30. Rours G, Verkooyen R, Willemse H, van der Zwaan E, van Belkum A, de Groot R, Verbrugh H, Ossewaarde J. Use of pooled urine samples and automated DNA isolation to achieve improved sensitivity and cost-effectiveness of large-scale testing for *Chlamydia trachomatis* in pregnant women. *Journal of Clinical Microbiology* 2005;43:4684–4690. [PubMed: 16145127]
31. Ornaghi J, March G, Boito G, Marinelli A, Beviacqua J, Giuggia J, Lenardon S. Infectivity in Natural Populations of *Delphacodes kuscheli* Vector of ‘Mal Rio Cuarto’ Virus. *Maydica* 1999;44:219223.
32. Shah D, Dillard H, Nault B. Sampling for the incidence of aphid-transmitted viruses in snap bean. *Phytopathology* 2005;95:1405–1411. [PubMed: 18943551]
33. Chen C, Swallow W. Using group testing to estimate a proportion, and to test the binomial model. *Biometrics* 1990;46:1035–1046. [PubMed: 2085624]
34. Thorburn D, Dundas D, McCrudden E, Cameron S, Goldberg D, Symington I, Kirk A, Mills P. A study of hepatitis C prevalence in healthcare workers in the west of Scotland. *Gut* 2001;48:116–120. [PubMed: 11115832]
35. Vansteelandt S, Goetghebeur E, Verstraeten T. Adjusting for confounding when estimating a time trend in HIV prevalence based on pooled serum samples. *Archives of Public Health* 1999;57:89–105.
36. Verstraeten T, Farah B, Duchateau L, Matu R. Pooling sera to reduce the cost of HIV surveillance: A feasibility study in a rural Kenyan district. *Tropical Medicine and International Health* 1998;3:747–750. [PubMed: 9754671]
37. Hosmer, D.; Lemeshow, S. *Applied Logistic Regression*. New York: Wiley; 2000.

Figure 1.

Model fits for individual and group testing for one simulated data set using different pooling strategies. The covariate is age. Corresponding model estimates are in Table 1.

Figure 2. Values of the slope estimate $\hat{\beta}_1$ in model (3) for individual and group testing using 100 simulated data sets with different pooling strategies.

Table 1

Parameter estimates, estimated asymptotic variance, and estimated asymptotic relative efficiency for one simulated data set; note that $\beta_0 = -1.97$ and $\beta_1 = -0.024$. Model fits are in Figure 1.

	Individual	AL	RA	DI
β_0	-1.45	-1.24	0.54	0.54
β_1	-0.042	-0.049	-0.130	-0.128
$\hat{\text{av}}(\hat{\beta}_1)$	0.0008	0.0011	0.0068	0.0293
$\hat{\text{re}}(\hat{\beta}_{I1} \text{ to } \hat{\beta}_{G1})$		0.70	0.12	0.03

Results using 100 simulated data sets with $\beta_0 = -1.97$ and $\beta_1 = -0.024$. Columns 2-4 list Pearson correlations for the 100 values of $\hat{\beta}_1$ Columns 5-8 list summary measures for the 100 values of $\hat{\beta}_1$. The 95 percent confidence interval is based on the t_{99} distribution.

Table 2

	Individual	AL	RA	Mean	Median	Variance	95% CI for mean
Individual	*	*	*	-0.025	-0.022	0.0007	(-0.030, -0.019)
AL	0.85	*	*	-0.025	-0.022	0.0010	(-0.032, -0.019)
RA	0.33	0.24	*	-0.039	-0.030	0.0071	(-0.056, -0.022)
DI	-0.05	-0.09	-0.13	-0.047	-0.055	0.0197	(-0.075, -0.019)

Table 3

Percent bias comparisons with a fixed number of individuals $I \times K$ for the model with $\beta_0 = -2, \beta_1 = 0.6931$, and $x_{ik} \sim \mathcal{U}(-7.079, 1.663)$. I denotes the common group size, and K denotes the number of groups.

I	$I \times K = 200$				$I \times K = 500$				$I \times K = 1000$			
	2	5	10	20	2	5	10	20	2	5	10	20
AL	10.8	20.3	30.7	9.1	3.5	5.5	13.9	29.7	1.4	2.5	5.8	15.7
RA	9.8	14.0	34.1	69.2	3.6	2.8	8.3	38.9	1.5	3.5	5.3	17.6
DI	7.9	9.6	20.6	121.3	2.8	7.8	20.4	42.4	0.7	3.0	9.3	37.7
Individual = 9.9												
Individual = 3.4												
Individual = 1.1												

Table 4

Average relative efficiency and correlation comparisons with a fixed number of individuals $I \times K$ for the model with $\beta_0 = -2$, $\beta_1 = 0.6931$, and $x_{ik} \sim \mathcal{U}(-7, 0.79, 1.663)$. I denotes the common group size, and K denotes the number of groups.

$\widehat{\text{are}}(\beta_{I1} \text{ to } \beta_{G1})$		$I \times K = 200$					$I \times K = 500$					$I \times K = 1000$				
I		2	5	10	20		2	5	10	20		2	5	10	20	
AL		0.87	0.76	0.57	0.27	0.93	0.93	0.78	0.59	0.33	0.94	0.94	0.79	0.59	0.33	
RA		0.64	0.29	0.12	0.04	0.68	0.68	0.30	0.12	0.04	0.69	0.69	0.31	0.13	0.04	
DI		0.46	0.08	0.02	0.01	0.49	0.49	0.08	0.02	0.01	0.50	0.50	0.09	0.02	0.01	
$\widehat{\text{corr}}(\beta_{I1}, \beta_{G1})$		$I \times K = 200$					$I \times K = 500$					$I \times K = 1000$				
I		2	5	10	20		2	5	10	20		2	5	10	20	
AL		0.97	0.79	0.56	0.30	0.97	0.97	0.89	0.65	0.47	0.97	0.97	0.88	0.72	0.46	
RA		0.85	0.56	0.32	0.15	0.85	0.85	0.60	0.35	0.21	0.82	0.82	0.35	0.36	0.17	
DI		0.71	0.30	0.09	0.08	0.73	0.73	0.37	0.12	0.01	0.73	0.73	0.33	0.16	0.09	

Table 5

Percent bias and relative efficiency comparisons with a fixed number of tests K for the model with $\beta_0 = -2, \beta_1 = 0.6931$, and $x_{ik} \sim \mathcal{U}(-7.079, 1.663)$. Values of $\widehat{\text{are}}(\widehat{\beta}_i \text{ to } \widehat{\beta}_{G1})$ larger than unity mean that group testing is more efficient than individual testing. I denotes the common group size, and K denotes the number of groups.

$\text{pbias}(\beta_1)$		$K = 100$				$K = 200$				$K = 500$			
I		2	5	10	2	5	10	2	5	10	2	5	10
AL		10.8	5.5	5.8	5.1	2.5	2.2	1.4	1.0	0.4			
RA		9.8	2.8	5.5	4.2	3.5	4.4	1.5	1.3	0.3			
DI		7.9	7.8	9.3	4.0	3.0	4.4	0.7	4.0	3.6			
Individual = 29.4													
$\widehat{\text{are}}(\beta_{I1} \text{ to } \beta_{G1})$													
Individual = 9.9													
$K = 200$													
I		2	5	10	2	5	10	2	5	10	2	5	10
AL		8.0	17.1	25.2	2.8	5.9	8.6	2.2	4.6	6.7			
RA		5.8	6.5	5.3	2.0	2.3	1.9	1.6	1.8	1.5			
DI		4.2	1.8	0.8	1.5	0.6	0.3	1.2	0.5	0.2			
$K = 500$													