Maximum likelihood estimators in regression models for error-prone group testing data

XIANZHENG HUANG, MD S. WARASI

Department of Statistics, University of South Carolina

ABSTRACT. Since the seminal work of Dorfman (1943) group testing has been widely adopted in epidemiological studies. In Dorfman's context of detecting syphilis in U.S. army recruits, group testing entails pooling recruits' blood samples and testing the pools, as opposed to testing individual samples. A negative pool indicates all individuals in the pool free of syphilis antigen, whereas a positive pool suggests at least one sample carry the antigen. With covariate information regularly collected along with disease status, many researchers have considered regression models that allow one study covariate-adjusted disease prevalence. In this article, we study maximum likelihood estimators of the covariate effects in these regression models when the group response is prone to error. We show that, when compared to inference drawn from individual testing data, inference based on group testing data can be more resilient to response misclassification in terms of both bias and efficiency. We also provide practically valuable guidance on designing the group composition to alleviate adverse effects of misclassification on inference results.

Key words: attenuation, efficiency, generalized linear model, individual testing.

1. Introduction

In the past few decades, group testing has received increasing attention in disease screening (Gastwirth and Hammick, 1989; Dhand, Johnson, Toribio, 2010), pollution detection (Wahed et al., 2006; Lennon, 2007), drug discovery (Remlinger et al., 2006), and genetics (Chi et al., 2009). One initial goal among statisticians in this line of research was to estimate the prevalence of a rare trait. This later evolved into the more challenging problem of estimating a covariate-adjusted prevalence (Vansteelandt, Goetghebeur, and Verstraeten, 2000; Xie, 2001). With covariates involved, regression analysis for group testing data has become a central interest (Chen, Tebbs, and Bilder, 2009; Delaigle and Meister, 2011; Delaigle and Hall, 2012; McMahan, Tebbs, and Bilder, 2013; Delaigle, Hall, and Wishart, 2014; Wang et al., 2014). Alongside the development of this methodology, McMahan, Tebbs, and Bilder (2012a,b) designed efficient group testing strategies that make use of covariate information.

From a practical standpoint, group testing is time/cost-efficient when the trait of interest is rare. From a statistical perspective, inferences based on group testing data typically suffer from lower efficiency since a positive group response is less informative than the collection of individual responses from that group. In this article, we rigorously show that, in the presence of misclassification in the binary response, one can actually benefit from drawing inference based on group testing data when compared to individual testing data.

Misclassification in group testing is common in many applications. For example, in the infertility prevention project (IPP), the state of Iowa uses group testing with the GenProbe Aptima Combo 2 Assay nucleic acid amplification test (Gen-Probe, San Diego) for chlamydia and gonorrhea, which is prone to error (Tebbs, McMahan, and Bilder, 2013). Viewing misclassification as a form of measurement error contamination, our findings reveal that likelihood-based inference from group testing data enjoys a certain level of robustness to error contamination in the response. This conclusion complements nicely the finding in Huang

and Tebbs (2009), which indicates that regression analysis based on group testing data can be more robust to measurement error in covariates when compared to regression analysis of individual testing data. Although mismeasured responses are as ubiquitous as error-prone covariates in practice, less research has examined regression analyses in the presence of the former, and no investigations have been pursued in the context of group testing. Our study fills in this important gap in the literature. In particular, by considering covariate-adjusted prevalence, this study generalizes the work of Hung and Swallow (1999), who showed gains in robustness when estimating disease prevalence of a homogeneous population. Viewing individual testing as a special case of group testing (with group size equal to one), our findings encompass those in Neuhaus (1999), who considered regression analysis based on error-prone individual testing responses.

To set the stage for our theoretical development regarding maximum likelihood estimators (MLE) of regression coefficients, we define notation and formulate models for group testing data in Section 2. In Section 3, we study asymptotic bias of the MLE when one ignores misclassification, leading to the so-called naive MLE. Then we study the effects of misclassification and grouping on the efficiency of the non-naive/consistent MLE obtained by accounting for misclassification in Section 4. In Section 5, we carry out regression analyses of two data sets, one from the IPP and one from a surveillance study of HIV in Kenya. Lastly, in Section 6, we summarize contributions of this paper and discuss further research topics. All appendices referenced henceforth are in the supplementary materials.

2. Data and models

Denote by Y_{ij} the true binary response of subject j in group i, and by X_{ij} the corresponding covariate, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. For ease of exposition, we consider a univariate covariate associated with each subject for the majority of the article. Generalization to multivariate covariates is mathematically straightforward, which is included in the sup-

plementary materials and demonstrated in an example in Section 5. Define the true group response as $Z_i = \max_{1\leq j\leq n_i} Y_{ij}$, and the $n_i \times 1$ vector of covariates as $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})^{\mathrm{T}}$, for $i = 1, \ldots, m$. Suppose that the model of Y_{ij} given X_{ij} is specified by the following generalized linear model (GLM),

$$P(Y_{ij} = 1|X_{ij}; \boldsymbol{\beta}) = g^{-1}(\beta_0 + \beta_1 X_{ij}), \text{ for } i = 1, \ldots, m, \; j = 1, \ldots, n_i, \tag{1}$$

where $g(\cdot)$ is a known, increasing, and differentiable link function, and $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ is the vector of regression coefficients, including the intercept β_0 and the covariate effect β_1. It follows that the model of Z_i conditional on $\mathbf{X}_i$ is determined by

$$P(Z_i = 1|\mathbf{X}_i; \boldsymbol{\beta}) = 1 - \prod_{j=1}^{n_i} \left\{1 - g^{-1}(\beta_0 + \beta_1 X_{ij})\right\}, \text{ for } i = 1, \ldots, m. \tag{2}$$

In this study, $\{Z_i\}_{i=1}^m$ are unobserved, and the observed group responses $\{Z_i^*\}_{i=1}^m$ result from potential misclassification of the true group responses. Suppose that the misclassification mechanism is dictated by the sensitivity, $\eta = P(Z_i^* = 1|Z_i = 1)$, and the specificity, $\theta = P(Z_i^* = 0|Z_i = 0)$, for $i = 1, \ldots, m$. Like in most measurement error problems where the severity of error contamination is unknown, validation data or external data are needed to infer η and θ (Hanson, Johnson, and Gastwirth, 2006; Küchenhoff, Mwalili, and Lesaffre, 2006). To focus on inference about $\boldsymbol{\beta}$, we do not introduce such data and we assume η and θ known. Moreover, we assume that η and θ do not depend on the group size n_i. This assumption is common in the group testing literature for single infections (Kim et al., 2007). For this to be reasonable in practice, assay thresholds may need to be adjusted to accommodate pooled and individual specimens (McMahan, Tebbs, and Bilder, 2013).

By (2), it is straightforward to show that the probability of observing a positive group response is, for $i = 1, \ldots, m$,

$$P_T(Z_i^* = 1|\mathbf{X}_i; \boldsymbol{\beta}) = \eta - (\eta + \theta - 1) \prod_{j=1}^{n_i} \left\{1 - g^{-1}(\beta_0 + \beta_1 X_{ij})\right\}, \tag{3}$$

where the subscript "T" in P_T signifies that (3) specifies the true model of Z_i^* given $\mathbf{X}_i$. If one ignores misclassification and views $\{Z_i^*\}_{i=1}^m$ to be the same as $\{Z_i\}_{i=1}^m$, one would carry out a naive regression analysis assuming the following wrong model of Z_i^* given $\mathbf{X}_i$,

$$P_F(Z_i^* = 1|\mathbf{X}_i;\boldsymbol{\beta}^*) = 1 - \prod_{j=1}^{n_i} \left\{1 - g^{-1}(\beta_0^* + \beta_1^* X_{ij})\right\}, \text{ for } i = 1, \ldots, m. \tag{4}$$

The subscript "F" in P_F is used to stress that (4) leads to a false model of Z_i^* given $\mathbf{X}_i$, and $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^{\mathrm{T}}$ is the vector of regression coefficients one would estimate under this false model, whose interpretation differs from that of $\boldsymbol{\beta}$ in (1)–(3).

3. Naive maximum likelihood estimator

3.1 *Estimating equations based on group testing data*

Denote by $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1)^{\mathrm{T}}$ the naive MLE of $\boldsymbol{\beta}$ that maximizes the observed-data likelihood derived from the wrong model in (4). By theories of maximum likelihood estimation based on a misspecified model (White, 1982), under regularity conditions, $\hat{\boldsymbol{\beta}}^*$ converges almost surely to $\boldsymbol{\beta}^*$ as $m \to \infty$, with $\max_{1\le i\le m} n_i$ bounded, where $\boldsymbol{\beta}^*$ minimizes the Kullback-Leibler (KL) divergence between the true observed-data likelihood, $f_T(\{Z_i^*, \mathbf{X}_i\}_{i=1}^m;\boldsymbol{\beta})$, and the false likelihood, $f_F(\{Z_i^*, \mathbf{X}_i\}_{i=1}^m;\boldsymbol{\beta}^*)$. More specifically, $\boldsymbol{\beta}^*$ minimizes

$$\begin{aligned}
\mathrm{KL}(f_T, f_F) &= \lim_{m\to\infty} m^{-1} E\left[\log \frac{f_T(\{Z_i^*, \mathbf{X}_i\}_{i=1}^m;\boldsymbol{\beta})}{f_F(\{Z_i^*, \mathbf{X}_i\}_{i=1}^m;\boldsymbol{\beta}^*)}\right] \\
&= \lim_{m\to\infty} m^{-1} \sum_{i=1}^m E_{\mathbf{X}_i}\left[E_{Z_i^*|\mathbf{X}_i}\left\{\log \frac{f_T(Z_i^*|\mathbf{X}_i;\boldsymbol{\beta})}{f_F(Z_i^*|\mathbf{X}_i;\boldsymbol{\beta}^*)}\right\}\right] \\
&= \mathcal{E}_{\mathbf{X}}\left[E_{Z_i^*|\mathbf{X}_i}\left\{\log \frac{f_T(Z_i^*|\mathbf{X}_i;\boldsymbol{\beta})}{f_F(Z_i^*|\mathbf{X}_i;\boldsymbol{\beta}^*)}\right\}\right],
\end{aligned} \tag{5}$$

where $f_T(Z_i^*|\mathbf{X}_i;\boldsymbol{\beta}) = P_T(Z_i^* = 1|\mathbf{X}_i;\boldsymbol{\beta})^{Z_i^*} P_T(Z_i^* = 0|\mathbf{X}_i;\boldsymbol{\beta})^{1-Z_i^*}$, $f_F(Z_i^*|\mathbf{X}_i;\boldsymbol{\beta}^*)$ is similarly defined, for $i = 1, \ldots, m$, and $\mathcal{E}_{\mathbf{X}}(\cdot)$ refers to $\lim_{m\to\infty} m^{-1}\sum_{i=1}^m E_{\mathbf{X}_i}(\cdot)$. The existence of a unique minimizer to (5) over $\boldsymbol{\beta}^*$ is guaranteed by the identifiability of $\boldsymbol{\beta}^*$ when testing errors are known.

In Appendix A we differentiate (5) with respect to $\boldsymbol{\beta}^*$ and show that $\boldsymbol{\beta}^*$ is the unique solution to the following system of estimating equations,

$$\mathcal{E}_{\mathbf{x}}\Big[\left\{\lambda(1|\mathbf{X}_i;\boldsymbol{\beta},\boldsymbol{\beta}^*)-\lambda(0|\mathbf{X}_i;\boldsymbol{\beta},\boldsymbol{\beta}^*)\right\}P_{F}(Z_i^*=0|\mathbf{X}_i;\boldsymbol{\beta}^*)\times \sum_{j=1}^{n}\frac{(1,\,X_{ij})^{\mathrm{T}}}{g'\left\{g^{-1}(\beta_0^*+\beta_1^*X_{ij})\right\}\left\{1-g^{-1}(\beta_0^*+\beta_1^*X_{ij})\right\}}\Big] \quad =\mathbf{0}_2, \tag{6}$$

where $g'(t)=(d/dt)g(t)$, $\mathbf{0}_2$ is the 2×1 vector of zeros, and

$$\lambda(z|\mathbf{X}_i;\boldsymbol{\beta},\boldsymbol{\beta}^*)=\frac{P_{T}(Z_i^*=z|\mathbf{X}_i;\boldsymbol{\beta})}{P_{F}(Z_i^*=z|\mathbf{X}_i;\boldsymbol{\beta}^*)},\quad \text{for } z=0,1. \tag{7}$$

The set of equations in (6) generally does not have an explicit solution, and one may numerically solve (6) for $\boldsymbol{\beta}^*$ given any true model/parameter configuration and pooling strategy. In the practice of group testing, typically one of the two pooling strategies, random pooling and homogeneous pooling, is employed. According to random pooling, pools are formed randomly and independently of covariate information. Homogeneous pooling requires one gather individuals with similar covariate values in a pool. Because the between-pool variability of covariate information is larger under homogeneous pooling than that under random pooling, more efficient statistical inference is expected based on data from homogeneous pooling (Vansteelandt, Goetghebeur, and Verstraeten, 2000).

3.2 *Approximate naive limiting maximum likelihood estimator*

To gain insight into the properties of $\boldsymbol{\beta}^*$, we first seek an approximation of $\boldsymbol{\beta}^*$, denoted by $\tilde{\boldsymbol{\beta}}^*$, in this subsection before we solve (6) numerically for $\boldsymbol{\beta}^*$ in the next subsection. The hope for $\tilde{\boldsymbol{\beta}}^*$ is that it is close to $\boldsymbol{\beta}^*$ enough such that it possesses some properties of $\boldsymbol{\beta}^*$ that reflect the influence of grouping and misclassification on $\boldsymbol{\beta}^*$. Detailed derivations leading to such $\tilde{\boldsymbol{\beta}}^*=(\tilde{\beta}_0,\tilde{\beta}_1^*)^{\mathrm{T}}$ are given in Appendix B. In what follows, we sketch the derivations leading to $\tilde{\beta}_1^*$, and discuss its implications.

The search for an approximated solution of (6) starts with assuming $n_i = n$, for $i = 1, \ldots, m$. We envision that $\tilde{\boldsymbol{\beta}}^*$ solves $\lambda(1|\mathbf{X}_i; \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}^*) = \lambda(0|\mathbf{X}_i; \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}^*)$ for all $\mathbf{X}_i$, which is equivalent to

$$\prod_{j=1}^{n} \left\{1 - g^{-1}(\tilde{\beta}_0^* + \tilde{\beta}_1^* X_{ij})\right\} = P_T(Z_i^* = 0|\mathbf{X}_i; \boldsymbol{\beta}), \quad \text{for all } \mathbf{X}_i. \tag{8}$$

This vision of $\tilde{\boldsymbol{\beta}}^*$ simplifies the problem of solving (6) to solving (8). Then, we evaluate (8) at two special covariate values, $\mathbf{X}_i = S_i \mathbf{1}_n$ and $\mathbf{X}_i = (S_i+1)\mathbf{1}_n$, where $\mathbf{1}_n$ is the $n \times 1$ vector of 1's, and S_i is some common covariate value shared by subjects in group i. These configurations of within-pool covariate values can be seen in a special homogeneous pooling, such as those considered in Farrington (1992) and Tu, Kowalski, and Jia (1999). As elaborated in Appendix B, following these two evaluations of (8), one can solve the resultant equations for $\tilde{\beta}_1^*$ and find that $\tilde{\beta}_1^* = H(\beta_1, n)$, with

$$H(\beta_1, n) = g\left(1 - \left[P_T\left\{Z_i^* = 0|\mathbf{X}_i = (S_i+1)\mathbf{1}_n; \boldsymbol{\beta}\right\}\right]^{1/n}\right) - g\left[1 - \left\{P_T(Z_i^* = 0|\mathbf{X}_i = S_i\mathbf{1}_n; \boldsymbol{\beta})\right\}^{1/n}\right]. \tag{9}$$

Close inspections of (9) provides several useful insights on the effects of grouping and misclassification on naive inference of the covariate effect. First, for a very large n, $\tilde{\beta}_1^*$ becomes very close to zero, although the trend is not necessarily monotone. In other words, with very big group sizes, one tends to conclude lack of association between the response and the covariate based on naive analysis. Second, when $\eta + \theta > 1$, (9) implies that $\tilde{\beta}_1^*$ has the same sign as that of β_1. Hence, the naive estimator of β_1 asymptotically conserves the sign of β_1. Third, if $\beta_1 = 0$, then $\tilde{\beta}_1^* = 0$, that is, $H(0, n) = 0$. Following this last remark, one has the first-order Taylor series approximation of $\tilde{\beta}_1^* = H(\beta_1, n)$ around $\beta_1 = 0$ as

$$\tilde{\beta}_1^* \approx H'(0, n)\beta_1, \tag{10}$$

where $H'(0, n)$ is defined as $(\partial/\partial\beta_1)H(\beta_1, n)$ evaluated at $\beta_1 = 0$, given by

$$H'(0, n) = \frac{(\eta + \theta - 1)p_{0Y}^{n-1} g'\left[1 - \left\{1 - \eta + (\eta + \theta - 1)p_{0Y}^n\right\}^{1/n}\right]}{\left\{1 - \eta + (\eta + \theta - 1)p_{0Y}^n\right\}^{1-1/n} g'(p_{1Y})}, \tag{11}$$

in which $p_{1Y} = g^{-1}(\beta_0)$ and $p_{0Y} = 1 - p_{1Y}$.

Straightforward algebraic manipulations reveal that setting $n = 1$ in $H'(0, n)$ in (11) results in the attenuation factor in Neuhaus (1999, equation (13)). In fact, with $\mathbf{X}_i = S_i\mathbf{1}_n$ for $i = 1, \ldots, m$, (2) reduces to $P(Z_i = 1|\mathbf{X}_i; \boldsymbol{\beta}) = 1-\{1-g^{-1}(\beta_0+\beta_1 S_i)\}^n$, for $i = 1, \ldots, m$, which is a GLM with link function $g^*(t) = g\{1 - (1 - t)^{1/n}\}$. This brings one back to the context of Neuhaus (1999). Hence, for this special case, one can reach (11) by replacing $g(\cdot)$ with $g^*(\cdot)$ in Neuhaus' attenuation result. Other than this special case where one can directly apply Neuhaus's bias analysis with some careful adjustments, our bias analysis encompasses a much broader scenario than that considered in the existing work. Despite the more complicated expression of $H'(0, n)$ and the generality of our data structure, we are able to establish the attenuation effect of misclassification on the naive MLE of β_1. This finding is stated in the next theorem, with the proof provided in Appendix C.

Theorem 3.1 If $\eta + \theta \geq 1$ and $1/g'(t)$ is concave, then $H'(0, n)$ in (11) lies in [0, 1] for all $n \geq 1$.

Theorem 3.1 indicates that $H'(0, n)$ can be interpreted as an attenuation factor associated with $\tilde{\beta}_1^*$ if $1/g'(t)$ is concave. The concavity of $1/g'(t)$ is assumed in the bias analysis in Neuhaus (1999) as well, where it is pointed out that any link function $g(t)$ defined as the inverse cumulative distribution function corresponding to a log-concave density function yields a concave $1/g'(t)$. In particular, popular link functions such as logistic, probit, and complementary log-log all share this characteristic.

To this end, one may wonder if the attenuation effect of misclassification on $\tilde{\beta}_1^*$ implied in Theorem 3.1 carries over to the exact naive limiting MLE, β_1^*. To address this, in the next subsection, we numerically solve (6) for $\boldsymbol{\beta}^*$ and compare β_1^* with (10). The upcoming large-sample numerical study (as opposed to finite-sample simulation studies) show that the approximation of $\tilde{\beta}_1^*$ given in (10) indeed captures certain features of β_1^* useful for under-

standing the effects of grouping and misclassification on the naive covariate effect estimator.

3.3 *Exact naive limiting maximum likelihood estimator*

In the numerical study, we set $\beta_1 = 1$ so that (10) becomes $\tilde{\beta}_1^* \approx H'(0, n)$. We use a logistic link in the GLM, $g(t) = \log\{t/(1-t)\}$, and consider $(\eta, \theta) = (0.9, 0.95), (0.9, 0.8)$, $\beta_0 = -3, -2.5, -2$. Covariate values are generated according to $X_{ij} \sim N(0, 1)$, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where $n_i = n$ $(i = 1, \ldots, m)$ with n varying from 1 to 15. Under this covariate configuration, the marginal probability $P(Y = 1)$ is around 0.069, 0.105, and 0.156 when $\beta_0 = -3, -2.5, -2$, respectively. These low prevalence rates are chosen to be consistent with most group testing applications, where typically a rare trait is of interest. We consider both homogeneous pooling and random pooling when solving (6), and denote by $\beta_{1,h}^*$ and $\beta_{1,r}^*$ the limiting MLE of β_1 from these two ways of pooling, respectively. To evaluate quantities in (6) under homogeneous pooling at a given level of n, we first generate a random sample of X of size $10^6 n$ from $N(0, 1)$; then we form 10^6 pools, each of size n, according to the sorted covariate values. Based on the simulated 10^6 homogeneous pools, we approximate the large-sample averages in (6), $\mathcal{E}_{\mathbf{x}}(\cdot)$, via the corresponding empirical means. The same approximation of the large-sample averages is employed when evaluating (6) for random pooling, but now the empirical means are computed using a random sample $\{\mathbf{X}_i, i = 1, \ldots, 10^6\}$ from $N(\mathbf{0}_n, \mathbf{I}_n)$, where $\mathbf{I}_n$ is the n-dimensional identity matrix.

Figure 1 depicts three asymptotic quantities, $\beta_{1,h}^*$, $\beta_{1,r}^*$, and $H'(0, n)$ (as an approximation of $\tilde{\beta}_1^*$) versus n. The pictorial comparison highlights the following three phenomena. First, more severe misclassification in the response leads to more attenuation in all considered naive MLEs. Second, when the prevalence is low, which is when group testing is most advocated, naive estimation of β_1 based on either pooling scheme can be less attenuated than that from individual testing data. Third, albeit we reach the analytic attenuation factor from an approximate solution to (6), $H'(0, n)$ preserves the overall trend of the two exact limiting

MLEs in terms of how they change as n increases.

The last point above suggests that $H'(0,n)$ is an informative indicator of the effect of misclassification on naive estimation that also reflects effects of grouping. With its explicit expression available, one may utilize (11) at the design stage of a group testing study to approximate an optimal group size in order to minimize the adverse effect of misclassification in regard to bias. In particular, we observe for a wide range of parameter configurations, besides those used to produce Figure 1, that the group size yielding the least amount of attenuation in the naive limiting MLE of β_1 when $\beta_1 \neq 0$ is always slightly below the group size that maximizes $H'(0,n)$.

Since ignoring misclassification means falsely setting η and θ at 1, one may wonder what if one misspecifies (η,θ) in various ways. Appendix D presents the limiting MLE of β_1 when one assumes sensitivity and/or specificity that deviate(s) from the truth in different ways. Although attenuation is not the universal phenomenon across all misspecified (η,θ), results there suggest that using group testing data does often produce less biased covariate effect estimator. Moreover, having the assumed testing errors closer to the truth typically yields less biased estimators. Hence, if one has a reliable estimates of η and θ, it is both natural and beneficial to incorporate them in inference rather than assuming them both to be one.

Certainly, the issue of bias is less of a concern, at least for large samples, if one conducts non-naive regression analysis using the correct (η,θ). This is the focus of the next section, where we study the efficiency of the resultant consistent MLE of $\boldsymbol{\beta}$.

4. Consistent maximum likelihood estimator

Denote by $\hat{\boldsymbol{\beta}}_{Z^*} = (\hat{\beta}_{0Z^*}, \hat{\beta}_{1Z^*})^{\mathrm{T}}$ the consistent MLE of $\boldsymbol{\beta}$ based on the correct model of Z^* in (3), and by $\hat{\boldsymbol{\beta}}_{Z} = (\hat{\beta}_{0Z}, \hat{\beta}_{1Z})^{\mathrm{T}}$ the counterpart consistent estimator in the absence of misclassification. In this section we study the asymptotic relative efficiency (ARE) of $\hat{\beta}_{1Z^*}$ to $\hat{\beta}_{1Z}$ defined as $\mathrm{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z}) = \mathrm{Var}(\hat{\beta}_{1Z})/\mathrm{Var}(\hat{\boldsymbol{\beta}}_{1Z^*})$, where each $\mathrm{Var}(\cdot)$ denotes

the asymptotic variance of the estimator derived from the Fisher information (Boos and Stefanski, 2013, Section 2.5) associated with the estimator. The primary interest here lies in the efficiency loss due to misclassification rather than the variance of a consistent estimator. This distinguishes our study from that in Liu et al. (2012), who studied the asymptotic variance of the estimated prevalence using error-prone group testing data, with no covariate involved.

4.1 *General efficiency results for group testing*

For $n_i \geq 1$, we show in detail in Appendix D that the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{Z^*}$ is

$$\text{Var}(\hat{\boldsymbol{\beta}}_{Z^*}) = \frac{1}{m}\begin{bmatrix} \mathcal{E}_{\mathbf{x}}(A^2C) & \mathcal{E}_{\mathbf{x}}(ABC) \\ \mathcal{E}_{\mathbf{x}}(ABC) & \mathcal{E}_{\mathbf{x}}(B^2C) \end{bmatrix}^{-1}, \tag{12}$$

where

$$A = \sum_{j=1}^{n_i} \frac{1}{(1-\mu_{Y_{ij}})g'(\mu_{Y_{ij}})},\ B = \sum_{j=1}^{n_i} \frac{X_{ij}}{(1-\mu_{Y_{ij}})g'(\mu_{Y_{ij}})},\ C = \frac{(\eta+\theta-1)^2(1-\mu_{Z_i})^2}{\mu_{Z_i^*}(1-\mu_{Z_i^*})}, \tag{13}$$

in which

$$\begin{aligned} \mu_{Y_{ij}} &= g^{-1}(\beta_0+\beta_1 X_{ij}), \quad \text{for } j=1,\ldots,n, \\ \mu_{Z_i} &= P(Z_i=1|\mathbf{X}_i;\boldsymbol{\beta}) = 1-\prod_{j=1}^{n_i}(1-\mu_{Y_{ij}}), \qquad (14) \\ \mu_{Z_i^*} &= P_T(Z_i^*=1|\mathbf{X}_i;\boldsymbol{\beta}) = \eta-(\eta+\theta-1)(1-\mu_{Z_i}). \qquad (15) \end{aligned}$$

And $\text{Var}(\hat{\boldsymbol{\beta}}_Z)$ follows immediately by substituting C in (12) with $C_1 = (1-\mu_{Z_i})/\mu_{Z_i}$, which is equal to C evaluated at $\eta=\theta=1$. By (12), the asymptotic variance of $\hat{\beta}_{1Z^*}$ and that of $\hat{\beta}_{1Z}$ are

$$\begin{aligned} \text{Var}(\hat{\beta}_{1Z^*}) &= \frac{1}{m}\left\{\mathcal{E}_{\mathbf{x}}(B^2C) - \frac{\mathcal{E}_{\mathbf{x}}(ABC)\mathcal{E}_{\mathbf{x}}(ABC)}{\mathcal{E}_{\mathbf{x}}(A^2C)}\right\}^{-1}, \qquad (16) \\ \text{Var}(\hat{\beta}_{1Z}) &= \frac{1}{m}\left\{\mathcal{E}_{\mathbf{x}}(B^2C_1) - \frac{\mathcal{E}_{\mathbf{x}}(ABC_1)\mathcal{E}_{\mathbf{x}}(ABC_1)}{\mathcal{E}_{\mathbf{x}}(A^2C_1)}\right\}^{-1}. \qquad (17) \end{aligned}$$

The ARE of $\hat{\beta}_{1Z^*}$ to $\hat{\beta}_{1Z}$, $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$, follows by forming the ratio of (17) over (16).

Setting $n_i = 1$ for $i = 1, \ldots, m$ in the above general results gives the asymptotic variance results for individual testing presented in Neuhaus (1999). There, the monotonicity of $\text{ARE}(\hat{\beta}_{1Y^*}, \hat{\beta}_{1Y})$ with respect to η and θ is explicitly established only for the special case with $\beta_1 = 0$, where $\hat{\beta}_{1Y^*}$ and $\hat{\beta}_{1Y}$ are the counterpart estimators of $\hat{\beta}_{1Z^*}$ and $\hat{\beta}_{1Z}$ for individual testing, respectively. Here, we are able to show in Appendix E the counterpart properties for $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ for an arbitrary β_1 and $n_i \geq 1$. These properties are summarized in Theorem 4.1.

Theorem 4.1 If $1 < \eta + \theta < 2$, then $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ obtained from (16) and (17) lies in (0, 1), and is increasing in η and θ for all β_1.

4.2 *Effects of grouping on efficiency*

With equal group size and $\beta_1 = 0$, $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ obtained in Section 4.1 reduces to

$$\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z}) = \frac{(\eta + \theta - 1)^2 p_{1Z} p_{0Z}}{p_{1Z^*} p_{0Z^*}}, \tag{18}$$

where p_{1Z} and p_{1Z^*} are equal to μ_Z and μ_{Z^*} in (14) and (15) evaluated at $\beta_1 = 0$, respectively, and $p_{0Z} = 1 - p_{1Z}$, $p_{0Z^*} = 1 - p_{1Z^*}$. We show in Appendix F that, with p_{1Y} and (η, θ) fixed, the ARE can be maximized at an n larger than 1. This finding is elaborated in Theorem 4.2, where we use $\text{Int}(t)$ to denote the integer that is closest to t.

Theorem 4.2 Define $p_{1Y} = g^{-1}(\beta_0) \in (0, 1)$ and $p_{0Y} = 1 - p_{1Y}$.

Case I: When $\eta = \theta \neq 1$, $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ in (18) is maximized at $n = 1$ if $p_{1Y} \geq 0.5$; and it is maximized at $n = \text{Int}(\log 0.5 / \log p_{0Y})$ if $p_{1Y} < 0.5$.

Case II: When $\eta \neq \theta \in (0.5, 1)$, $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ in (18) is maximized at

$$n = \max\left(1, \text{Int}\left[\log \frac{\eta(1-\eta) - \{\eta(1-\eta)\theta(1-\theta)\}^{1/2}}{(\eta + \theta - 1)(\theta - \eta)} \Big/ \log p_{0Y}\right]\right). \tag{19}$$

When $\beta_1 \neq 0$, one can numerically obtain the ARE associated with the covariate effect of

interest given any model/parameter/covariate configurations based on (16) and (17). Adopting the settings of the large-sample study in Section 3.3, we compute $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ when $\beta_1 = 0, 0.5, 1$. Under the current configurations, a larger β_0 or β_1 gives a higher prevalence rate, $P(Y = 1)$, which ranges from 0.047 to 0.155 here. Figure 2 presents these asymptotic quantities, which convey two key messages. First, $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ decreases as misclassification happens more often. Second, when the prevalence rate is low, non-naive inference based on group testing data can be more robust to misclassification in terms of efficiency than inference drawn from individual testing data. Moreover, the value of n that maximizes $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ when $\beta_1 = 0$ coincides with what is indicated by Theorem 4.2.

It is worth noticing in Figure 2 that many features of the ARE when $\beta_1 = 0$ are also observed when $\beta_1 \neq 0$, and the overall trend of how the ARE in (18) varies as n varies can be a reasonable approximation of the phenomenon when $\beta_1 \neq 0$. We observe that, in general, the group size maximizing the ARE when $\beta_1 \neq 0$ is smaller than the corresponding optimal group size when $\beta_1 = 0$. Therefore the optimal group size given in Theorem 4.2 can serve as an upper bound of a preferable group size when planning for a group testing study.

5. Applications

Here we consider two studies where group testing is involved. Depending on the type of data available in each study, we focus on the attenuation effect of misclassification on naive MLEs in the first study, and look into the efficiency loss in the consistent MLEs due to misclassification in the second study.

5.1 *IPP data for chlamydia*

As described in Section 1, subjects enrolled in the IPP are tested using an imperfect test for chlamydia and/or gonorrhea. Besides the testing results, covariate information of subjects, such as age, race, ethnicity, are also recorded. These covariate information are

potentially useful for predicting prevalence and designing efficient group testing strategies (McMahan, Tebbs, and Bilder, 2012a,b). In order to compare inference from individual testing data with inference from group testing data, we opt to use data from the state of Nebraska, where individual testing data are collected. In particular, we will analyze data regarding chlamydia from year 2008 for 14,441 female subjects, whose specimens were collected via cervix swabs. Also, to illustrate individual/group-testing regression with multivariate covariates, we include age ($X_{ij,1}$) and race ($X_{ij,2}$) as covariates of interest, and denote by $\boldsymbol{\beta}_1 = (\beta_{1,1}, \beta_{1,2})^{\mathrm{T}}$ the vector of corresponding covariate effects.

In this sample, 12% of the females were under the age of 18 and 80% between age 18 to 29. Caucasians (for which $X_{ij,2} = 1$) contribute 81% of this sample, and African-Americans (for which $X_{ij,2} = 2$) account for 11%. Finally, 7.19% were tested positive for chlamydia in this sample of size 14,441. With the response being the indicator of a positive test for chlamydia, considering a logistic regression for the individual testing data, we carry out both naive regression analysis (by assuming a perfect test) and non-naive regression analysis. To perform the non-naive analysis, we adopt the sensitivity $\eta = 0.942$ and specificity $\theta = 0.976$ given in Tebbs, McMahan, and Bilder (2013, Table 2). After this round of individual-testing data analysis, we create six group testing data sets resulting from the combination of two pooling strategies (homogeneous pooling and random pooling) and three group-size configurations. According to the first configuration, 2063 pools of equal size are created, with $n_i = 7$, for $i = 1, \ldots, 2063$. Under the second configuration, we form a total of 963 pools, with $n_i = 15$, for $i = 1, \ldots, 962$ and $n_{963} = 11$. The third configuration leads to pools of unequal sizes, with 509 pools of size 5, 1000 pools of size 7, and 544 pools of size 9. When homogeneous pooling is employed, before partitioning the raw individual testing data into pools, we sort the data first by age then by race. Based on each of the six sets of group testing data, we implement naive analysis and non-naive analysis successively. Table 1 presents these

estimates of the covariates effects from the individual testing data and six induced group testing data sets. Also listed in Table 1 is an estimate of the attenuation factor, $\hat{H}'$, defined as the ratio of the naive estimate over its non-naive counterpart. Appendix G provides the `SAS IML` code for computing the individual-testing naive estimates, group-testing naive estimates for both random pooling and homogeneous pooling under the second pooling configuration, and the non-naive counterpart estimates of all the above.

The consistent estimates of $\boldsymbol{\beta}_1$ from most of the above analyses seem to suggest significant covariate effects for both age and race. The naive estimates of $\boldsymbol{\beta}_1$ from individual testing data suffers far more attenuation due to misclassification than the naive estimates from any one of the six group testing data sets. This is reminiscent of the bias analysis of naive MLEs in Section 3. Moreover, we evaluate the attenuation factor in (11) at a range of n in the context of the data analyzed here, and find that the attenuation factor is maximized at $n = 8$ in the current setting. Suppose one wishes to use group testing in Nebraska in the follow-up study using a new assay, of which one is unsure about the testing errors. Following the discussion regarding a preferable group size to alleviate attenuation effect in Section 3.3, we would recommend a group size slightly below 8 to avoid too much underestimation of covariate effects due to inaccurate testing errors used in the likelihood-based inference.

5.2 *Kenyan data for HIV*

The data set to be entertained in this subsection was analyzed in Vansteelandt, Goetghebeur, and Verstraeten (2000), who compared the cost and precision of parameter estimates when individual testing data were used with the estimates from group testing data created according to different pooling strategies. This data set was collected during the first year of a surveillance study in Kenya aimed at assessing the HIV epidemic at that time and monitoring progress in the following years. More specifically, during that year of the study, individual testing data were collected in order to evaluate the feasibility of using group test-

ing data in the upcoming years for estimating covariate-adjusted HIV prevalence, where the recorded covariates include parity, age, education, etc. The binary response for this study is defined as the indicator of being HIV positive suggested by tests on subjects' serum samples. For illustration purposes, we consider parity as the covariate. Similar to the treatment on missing data in Vansteelandt, Goetghebeur, and Verstraeten (2000), we only use data from the 705 adult females, out of a total of 787, whose information of parity and test results were not missing. Among these 705 subjects, the values of parity range from 0 to 12, with 95% of them below 6. Because the sensitivity and specificity of the test were reported to be 1 and 0.9997, respectively, we treat the observed responses error-free. With a (nearly) perfect test, we can artificially create error-prone individual testing responses and group testing responses, which allows us to observe the efficiency loss due to misclassification in the non-naive inference.

Before we create artificial error-prone responses, with the logistic link and parity as the covariate in (1), we first compute MLEs of the regression parameters using the observed (deemed error-free) individual testing data of 705 subjects, denoted by $\hat{\boldsymbol{\beta}}_Y = (\hat{\beta}_{0Y}, \hat{\beta}_{1Y})^{\mathrm{T}}$. Then we create an induced group testing data set via random pooling with $n_i = 3$ for $i = 1, \ldots, 235$, based on which we obtain the MLEs of $\boldsymbol{\beta}$, namely, $\hat{\boldsymbol{\beta}}_Z$. The estimates of the covariate effect are $\hat{\beta}_{1Y} = -0.2488$ (0.0930) and $\hat{\boldsymbol{\beta}}_{1Z} = -0.2633$ (0.1801), with the corresponding estimated standard errors in parentheses, obtained following the sandwich variance estimation (Stefanski and Boos, 2002). Next, we create error-prone responses based on the observed individual testing outcomes using a sensitivity and a specificity of 0.98. Finally, we compute two sets of non-naive MLEs (by acknowledging $\eta = \theta = 0.98$), first using the error-prone individual testing data, resulting in estimates denoted by $\hat{\boldsymbol{\beta}}_{Y^*} = (\hat{\beta}_{0Y^*}, \hat{\beta}_{1Y^*})^{\mathrm{T}}$, second $\hat{\boldsymbol{\beta}}_{Z^*}$ based on an induced group testing set resulting from random pooling like previously done. The estimates of the covariate effect are $\hat{\beta}_{1Y^*} = -0.2550$ (0.1407) and $\hat{\beta}_{1Z^*} = -0.2721$

(0.1862). Comparing this round of estimation with the previous round before contaminating the observed responses, we have an estimate of $\text{ARE}(\hat{\beta}_{1Y^*}, \hat{\beta}_{1Y})$ and an estimate of $\text{ARE}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z})$ by forming the ratios of the corresponding estimated variances. These estimates are $\hat{\text{ARE}}(\hat{\beta}_{1Y^*}, \hat{\beta}_{1Y}) \approx 0.4368$ and $\hat{\text{ARE}}(\hat{\beta}_{1Z^*}, \hat{\beta}_{1Z}) \approx 0.9356$. This reinforces the finding in Section 4 that efficiency loss in consistent estimators due to misclassification can be less for group testing data than for individual testing data when the trait of interest is rare. Among the 705 individuals in this study, around 7.2% were found HIV positive.

If one wishes to plan in the future a group testing study for a comparable population, assuming $\eta = \theta \neq 1$ and using 0.072 as an estimate of p_{1Y}, one may apply Theorem 4.2 and find that, if $\beta_1 = 0$, the optimal group size is $n = 9$. Because the analysis using the error-free individual testing data seems to suggest $\beta_1 \neq 0$, one may consider a group size smaller than 9 to protect against too much efficiency loss in the consistent maximum likelihood estimation due to misclassification.

6. Discussion

With the individual testing response being a special case of the group testing response (with group size always equal to 1), our study leads to some interesting discoveries regarding the comparisons between group-testing inference and individual-testing inference. Although statistical inference in the presence of misclassification are understandably compromised by misclassification in responses, whether it is grouped responses or individual responses, we show that one can gain from group-testing inference more robustness to misclassification in regard to both accuracy and precision. Although robustness here does not imply consistency or no loss of efficiency, it is comforting for group testing practitioners that, compared to inference from a more costly study that produces individual testing data, inference based on the cheaper group testing data can be less compromised by misclassification.

The contribution of this study is at least twofold. First, the robustness features estab-

lished in this study elongate the list of advantages of group testing compared to individual testing. This is a valuable addition especially considering that group testing has been mostly advocated for its savings in time and cost, rarely for its gain in statistical inference. Second, as a special form of error contamination, misclassification in binary responses has been studied far less than problems involving error-prone covariates. There is the folklore in the field of measurement error implying that measurement error compromise statistical inference, whether one ignores it or accounts for it when making inference. Although we do not disagree with the folklore, we show in this study that the adverse effects can be alleviated if error-prone binary responses are strategically pooled in groups. Our theoretical findings provide practical guidance on such strategic pooling. To recap, if one ignores misclassification in inference, one should use a group size slightly below that maximizes the attenuation factor in (11) to avoid too much bias; if one accounts for the testing errors, one may use the group size no larger than that specified in Theorem 4.2 to alleviate efficiency lost.

Stepping outside of the above two research fields, results from this study have further intriguing implications. Even without misclassification, certainly efficiency loss is inevitable when group testing data are used for likelihood-based inference in place of individual testing data. Even without pooling individuals into groups, expectedly inference becomes biased in the presence of misclassification and it is uncounted for, except in special cases such as when $\beta_1 = 0$. Both actions of misclassifying and pooling lead to coarsened data, that is, data with less information than before the action takes place. Effects of each form of coarsening by itself on statistical inference are better understood, and mostly agree with one's intuition. But when both forms of coarsening occur, the interaction effects are far less intuitive and merit more careful theoretical investigation. Our findings in this study suggest that, as undesirable as each coarsening is from the statistical standpoint, the interplay of two coarsening can unexpectedly yield inference outcomes with some desirable properties.

This implication points at the bigger picture of inference based on coarsened data, and, in particular, raises the question of when it is possible, and how, to gain more from less informative data.

Supporting information

Additional supporting information for this article is available online:

References

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference, Theory and Methods*. Springer, New York.

Chen, P., Tebbs, J. M., & Bilder, C. R. (2009). Group testing regression models with fixed and random effects. *Biometrics*, 65, 1270–1278.

Chi, X., Lou, X., Yang, M., and Shu, Q. (2009). An optimal DNA pooling strategy for progressive fine mapping. *Genetica*, 135, 267–281

Delaigle, A. & Hall, P. (2012). Nonparametric regression with homogeneous group testing data. *The Annals of Statistics*, 40, 131–158.

Dhand, N. K., Johnson, W. O., & Toribio, J. L. (2010). A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 452–473.

Delaigle, A., Hall, P., & Wishart J. R. (2014). New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data. *Biometrika*, 567–585.

Delaigle, A. & Meister, A. (2011). Nonparametric regression analysis for group testing. *Journal of the American Statistical Association*, 106, 640–650.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14, 436–440.

Farrington, C. P. (1992). Estimating prevalence by group testing using generalized linear models. *Statistics in Medicine*, 11, 1591-1597.

Gastwirth, J. L. & Hammick, P. A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Applications to estimating the prevalence of AIDS antibodies in blood donors. *Journal of Statistical Planning and Inference*, 22, 15–27.

Hanson, T. E., Johnson, W. O., & Gastwirth, J. L. (2006). 'Bayesian inference for prevalence and diagnostic test accuracy based on dual-pooled screening. *Biostatistics*, 7, 41–57.

Huang, X. & Tebbs, J. M. (2009). On latent-variable model misspecification in structural measurement error models for binary response. *Biometrics*, 65, 710–718.

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

Küchenhoff, H., Mwalili, S., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, 62, 85–96.

Kim, H. Y, Hudgens, M. G., Dreyfuss, J. M., Westreich, D. J., & Pilcher, C. D. (2007). Comparison of group testing algorithms for case identification in the presence of testing error. *Biometrics*, 63, 1152–1163.

Lennon, J. T. (2007). Diversity and metabolism of marine bacteria cultivated on dissolved DNA. *Applied and Environmental Microbiology*, 73, 2799–2805.

Liu, A., Liu, C., Zhang, Z., & Albert, P. S. (2012). Optimality of group testing in the presence of misclassification. *Biometrika*, 99, 245–251.

McMahan, C. S., Tebbs, J. M., & Bilder, C. R. (2012a). Informative Dorfman Screening. *Biometrics*, 68, 287–296.

McMahan, C. S., Tebbs, J. M., & Bilder, C. R. (2012b). Two-dimensional informative array testing. *Biometrics*, 68, 793–804.

McMahan, C. S., Tebbs, J. M., & Bilder, C. R. (2013). Regression models for group testing data with pool dilution effects. *Biostatistics*, 14, 284–298.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86, 843–855.

Remlinger, K. S., Young, S. S., Hughes-Oliver, J. M., & Lam R. L. (2006). Statistical design of pools using optimal coverage and minimal collision. *Technometrics*, 48, 133–143.

Stefanski, L. A. & Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56, 29–38.

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

Tu, X. M., Kowalski, J., & Jia, G. (1999). Bayesian analysis for prevalence with covariates using simulation-based techniques: Applications to HIV screening. *Statistics in Medicine*, 18, 3059-3073.

Vansteelandt, S., Goetghebeur, E., & Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum sample. *Biometrics*, 56, 1126–1133.

Wahed, M. A., Chowdhury, D., Nermell, B., Khan, S. I., Ilias, M., Rahman, M., Persson, L. A., & Vahter, M. (2006). A modified routine analysis of arsenic content in drinking-water in Bangladesh by hydride generation-atomic absorption spectrophotometry. *Journal of Health, Population and Nutrition*, 24, 36–41.

Wang, D., McMahan, C. S., Gallagher, C. M., & Kulasekera K. B. (2014). Semiparametric group testing regression models. *Biometrika*, 101, 587–598.

White, H. L. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

Xie, M. (2001). Regression analysis on group testing samples. *Statistics in Medicine*, 20, 1957–1969.

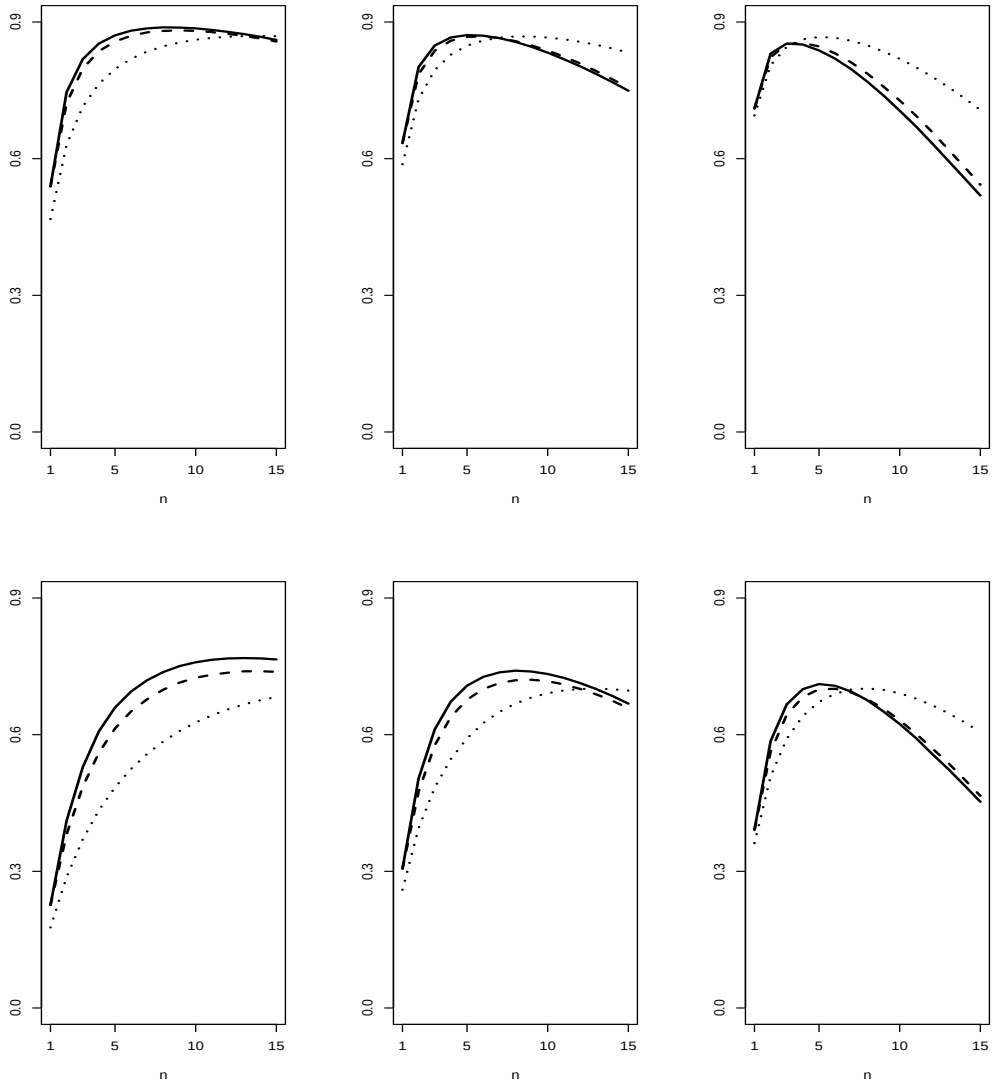

Fig. 1. The limiting MLE of β_1 under homogeneous pooling, $\beta^*_{1,h}$ (solid lines), the limiting MLE under random pooling, $\beta^*_{1,r}$ (dashed lines), and the attenuation factor $H'(0,n)$ (dotted lines) versus group size n when the prevalence rate is approximately 0.069 (left panels), 0.105 (middle panels), and 0.156 (right panels), with $(\eta,\theta) = (0.9,\ 0.95)$ (upper panels), and (0.9, 0.8) (lower panels).

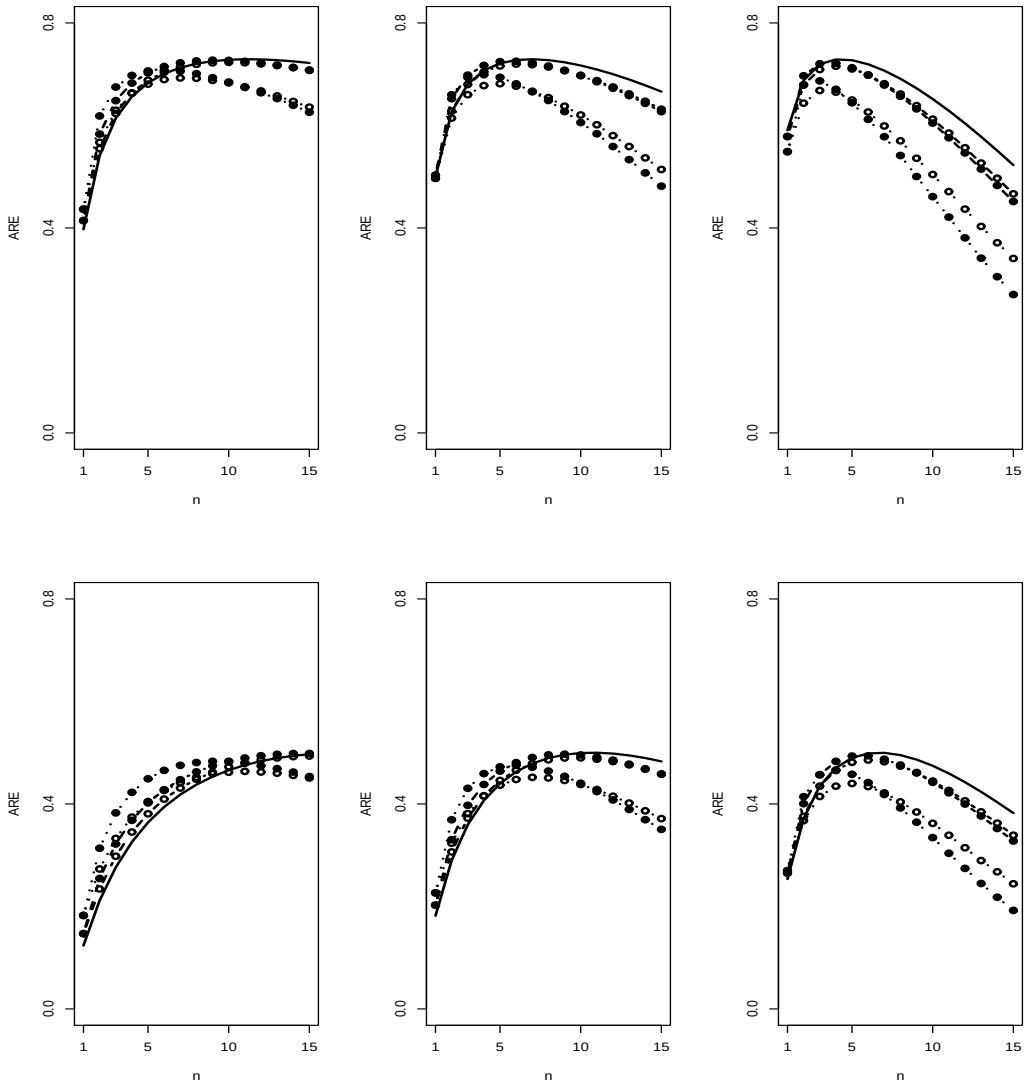

Fig. 2. Asymptotic relative efficiency (ARE) of $\hat{\beta}_{1Z^*}$ to $\hat{\beta}_{1Z}$ versus group size n when $\beta_0 = -3$ (left panels), -2.5 (middle panels), -2 (right panels), with $(\eta, \theta) = (0.9, 0.95)$ (upper panels), $(0.9, 0.8)$ (lower panels). Each panel includes ARE when $\beta_1 = 0$ (solid lines), 0.5 (dashed lines passing solid or empty circles), and 1 (dotted lines passing solid or empty circles). When $\beta_1 \neq 0$, ARE associated with homogeneous pooling are depicted by lines passing solid circles, and ARE associated with random pooling are depicted by lines passing empty circles.

Table 1

Naive maximum likelihood estimates and consistent maximum likelihood estimates of $\boldsymbol{\beta}_1$ computed using IPP data, where $\boldsymbol{\beta}_1$ contains two elements, $\beta_{1,1}$ and $\beta_{1,2}$, associated with age and race effects, respectively. Attenuation factors are estimated by $\hat{H}'$. Numbers in parentheses beneath the estimates are estimated standard errors of the estimates based on sandwich variance estimation (Stefanski and Boos, 2002). The code "UE" in the first column refers to the case with the third group-size configuration

	Naive estimates		Non-naive estimates		$\hat{H}'$	
	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,1}$	$\beta_{1,2}$
Individual testing	−0.0632	0.0713	−0.0998	0.0908	0.6325	0.7857
	(0.0081)	(0.0178)	(0.0119)	(0.0219)		
Homogeneous pooling						
$n = 7$	−0.0550	0.0643	−0.0606	0.0687	0.9081	0.9368
	(0.0081)	(0.0228)	(0.0091)	(0.0248)		
$n = 15$	−0.0439	0.0342	−0.0498	0.0389	0.8819	0.8792
	(0.0089)	(0.0314)	(0.0105)	(0.0370)		
UE	−0.0555	0.0670	−0.0576	0.0725	0.9620	0.9244
	(0.0080)	(0.0235)	(0.0072)	(0.0205)		
Random pooling						
$n = 7$	−0.0851	0.0811	−0.0911	0.0864	0.9341	0.9396
	(0.0186)	(0.0363)	(0.0203)	(0.0380)		
$n = 15$	−0.0996	0.1028	−0.1104	0.1112	0.9020	0.9241
	(0.0314)	(0.0589)	(0.0352)	(0.0634)		
UE	−0.0774	0.0891	−0.0822	0.0932	0.9408	0.9565
	(0.0200)	(0.0370)	(0.0214)	(0.0392)		