Received: 3 June 2018 | Revised: 3 March 2019 | Accepted: 14 June 2019

DOI: 10.1002/sim.8311

RESEARCH ARTICLE

WILEY Statistics in Medicine

Regression analysis and variable selection for two-stage multiple-infection group testing data

Juexin Lin[1] | **Dewei Wang**[1] | **Qi Zheng**[2]

[1]Department of Statistics, University of South Carolina, South Carolina,

[2]Department of Bioinformatics and Biostatistics, University of Louisville, Kentucky,

Correspondence
Dewei Wang, Department of Statistics, University of South Carolina, Columbia, SC 29028.
Email: deweiwang@stat.sc.edu

Present Address
Dewei Wang, Department of Statistics, University of South Carolina, Columbia, SC 29028

Funding information
National Institutes of Health, Grant/Award Number: R03 AI135614

Group testing, as a cost-effective strategy, has been widely used to perform large-scale screening for rare infections. Recently, the use of multiplex assays has transformed the goal of group testing from detecting a single disease to diagnosing multiple infections simultaneously. Existing research on multiple-infection group testing data either exclude individual covariate information or ignore possible retests on suspicious individuals. To incorporate both, we propose a new regression model. This new model allows us to perform a regression analysis for each infection using multiple-infection group testing data. Furthermore, we introduce an efficient variable selection method to reveal truly relevant risk factors for each disease. Our methodology also allows for the estimation of the assay sensitivity and specificity when they are unknown. We examine the finite sample performance of our method through extensive simulation studies and apply it to a chlamydia and gonorrhea screening data set to illustrate its practical usefulness.

KEYWORDS

adaptive LASSO, multiplex assay, pooled testing, sensitivity, specificity

1 | INTRODUCTION

1.1 | Motivation

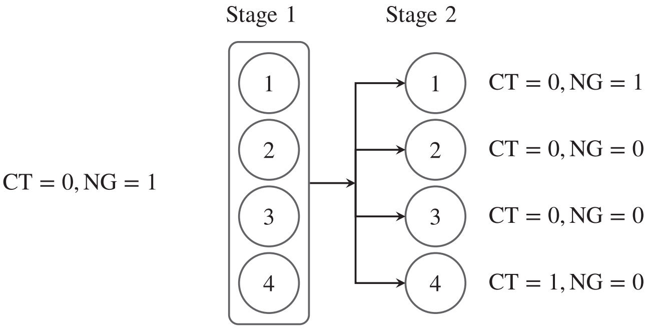
This article is motivated by the annual *Chlamydia trachomatis* (CT) and *Neisseria gonorrhoeae* (NG) screening practice conducted by the State Hygienic Laboratory (SHL) in Iowa. The CT and NG are two of the most common notifiable sexually transmitted diseases (STDs) in the United States. Over two million cases were reported to the Centers for Disease Control and Prevention (CDC) in 2016.[1] Both infections are commonly asymptomatic in women. If left untreated, they could cause pelvic inflammatory disease and further lead to tubal infertility, ectopic pregnancy, or chronic pelvic pain.[2] In addition, both diseases could facilitate the transmission of human immunodeficiency virus and human papillomavirus infection.[3] Concerned by these severe sequelae, CDC continually supports nationwide CT/NG screening and recommends annual CT/NG screening for all sexually active women under 25 years old.[4]

In this nationwide screening practice, specimens (swab or urine) are collected across each state and shipped to major state laboratories to be tested. Due to different budgets, laboratories conduct the screening differently. For example, the Nebraska Public Health Laboratory (NPHL) uses a traditional individual testing protocol that tests individual specimens one by one. The SHL tests male specimens and female urine specimens individually, but it tests female swab specimens according to a two-stage pooling protocol:

The SHL pooling protocol

- Individual swab specimens are randomly assigned to nonoverlapping groups of size four. A pool is constructed by mixing individual specimens in the same group.

FIGURE 1 A possible set of the State Hygienic Laboratory (SHL) pooled testing outcomes from a group of four individuals: The rectangle with rounded corners represents the pooled specimen that is constructed by mixing the four individual specimens (in circles) together. The pool tested negative for *Chlamydia trachomatis* (CT) (ie, CT = 0) but positive for *Neisseria gonorrhoeae* (NG) (ie, NG = 1). As per the SHL pooling protocol, the positivity of NG triggered the second stage of screening for both infections. Due to testing errors, we see a discrepancy between the two stages, ie, the fourth individual retested positive for CT but the pool tested negative for CT



- **Stage 1**. Each pool is tested for CT and NG simultaneously using a multiplex assay. If a pool tests negative for both infections, all the involved individuals are diagnosed as negative for each infection with no additional tests; otherwise, the protocol proceeds to the next stage.
- **Stage 2**. Swabs of individuals in pools that test positive for either infection are retested separately using the same multiplex assay for final diagnosis.

The most practical reason of using pooling is cost reduction. When a pool tests negative for both infections, four individuals are diagnosed at the expense of one assay. Since switching from individual testing to pooling in 1999, Iowa has saved over \$2.2 million in the CT/NG screening.[5]

As per the screening guidelines, many risk factors are collected as well, such as age, number of partners, any symptoms of the infections, etc. A motivating question is how to incorporate this covariate information so that one can identify truly relevant risk factors for each infection and understand their effects. Challenges to this question arise from the use of the multiplex Aptima Combo 2 Assay (Gen-Prob Inc, San Diego, CA), an imperfect discriminatory test that produces diagnoses for both diseases simultaneously. Due to the imperfectness of the assay, it is possible to observe some discrepancies between testing outcomes of the two stages, as shown in Figure 1. Whenever a discrepancy occurs, the SHL ignores pooled-level results from Stage 1 and makes the diagnosis solely based on individual testing from Stage 2. However, when the objective is probing the impact of risk factors rather than case identification, disregarding testing outcomes from any stage could impair the estimation. It is important to seamlessly incorporate outcomes from both stages. Towards this goal, we need to account for how likely the retests were triggered by either infection.

1.2 | Literature review

Pooled testing (also known as group testing) was initially proposed to screen for syphilis among War World II American army recruits.[6] Since this seminal work, pooling techniques have been successfully implemented to screen for many other infectious diseases, including HIV, hepatitis B virus (HBV), hepatitis C virus (HCV),[7] influenza,[8] and herpes.[9] Besides disease screening, many other areas, including genetics,[10] veterinary science,[11] medical entomology,[12] blood safety,[13] and drug discovery,[14] have also used the method of pooling. Statistical research in group testing primarily focused on improving the diagnostic accuracy and cost-saving ability of a pooling protocol[15] or estimating individual-level characteristics from pooled testing data. This article falls into the latter category. When group testing data only involves a single infection, the research focusing on estimation started with estimating a disease prevalence.[16,17] This research avenue was expanded to incorporate individual covariate information through the use of parametric regression models, such as generalized linear models,[18,19] mixed models,[20] and Bayesian regression models.[21] Semiparametric and nonparametric regression methods have also been developed.[22,23] However, all these works are limited to one infection.

The use of multiplex assays makes pooled testing data with multiple infections widely available. For example, in addition to CT/NG, HIV/Syphilis, HIV/HCV, or HIV/HBV/HCV can be detected simultaneously.[24] In statistical literature, the research focusing on estimation with multiple-infection group testing data is scarce. A few works have studied the estimation of disease prevalence.[25-28] Regression analysis for this type of data remains mostly untapped. To the best of our knowledge, the only work is an approach based on generalized estimating equations.[29] However, it did not consider retesting outcomes arising from the second stage of screening and thus does not apply to the SHL screening practice.

When using an imperfect assay, the values of assay sensitivity and specificity are crucial to estimation in pooled testing. Most of the aforementioned literature assumed that there were some preliminary studies to provide those misclassification parameters. However, this assumption could be impractical because the preliminary study might have used unrepresentative samples.[17] If inaccurate values of assay sensitivity and specificity were used for estimation, it could compromise inference. In this article, we keep the testing error rates as unknown and estimate them from the data along with the regression.

Existing literature has not considered the combination of incorporating retesting results into regression and estimating misclassification parameters in the context of multiple-infection group testing. Only one Bayesian work has provided inference for disease prevalence and estimates of assay sensitivity and specificity without consideration of individual covariates.[27] In this article, we propose a copula-based multivariate binary regression model to incorporate the covariates. We introduce a generalized expectation-maximization (GEM) algorithm to facilitate the numerical computation of the maximum likelihood estimates (MLEs) of the regression coefficients and misclassification parameters. When compared with the traditional expectation-maximization algorithm, the GEM only requires the maximization step to search for an increase in the objective function rather than achieving the maximum.[30,31] This feature greatly accelerates the computation of the MLE.

In addition, we provide a variable selection technique that can identify truly relevant risk factors for each infection. A recent work has introduced a regularized regression technique for group testing.[32] However, it is for a single infection. Our work is designed to allow for multiple infections. We believe a package of regression, estimation of misclassification parameters, and variable selection can provide a useful toolbox for the epidemiology study of CT and NG based on group testing data.

The rest of the article is organized as follows. In Section 2, we propose a new copula-based regression model for multiple-infection group testing data. In Section 3, we introduce the GEM algorithm that accelerates the computation of the MLE. Section 4 presents a variable selection method that can identify important risk factors for each infection. In Section 5.1, we use simulation to illustrate that, with the use of a fewer number of tests, the SHL pooling protocol can lead to more efficient regression estimates, better prediction of infection probabilities, and more accurate variable selection than traditional individual testing. These advantages are further demonstrated by analyzing a CT/NG screening data set in Section 5.2. Section 6 presents a discussion of this work. All technical details and additional numerical results are relegated to the supplementary materials.

2 | MODEL

Suppose N individuals are to be tested. We randomly assign each individual to one of J groups, each of size c_j, ie, $N = \sum_{j=1}^{J} c_j$. For generality, we allow group size c_j to vary across groups. Motivated by the CT/NG screening practice, we mainly consider two infections. Section 6 discusses an extension of more than two diseases. The true infection statuses of the ith individual in the jth group are denoted by a binary vector $\widetilde{\boldsymbol{Y}}_{ij} = (\widetilde{Y}_{ij1}, \widetilde{Y}_{ij2})^{\mathrm{T}}$, where $\widetilde{Y}_{ijk} = 1$ if the individual is positive for the kth infection and $\widetilde{Y}_{ijk} = 0$ if otherwise, for $i = 1, \ldots, c_j, j = 1, \ldots, J$, and $k = 1, 2$. Denote the covariates (risk factors and an intercept term) of the ith individual in the jth group by a $(p+1)$-dimensional vector $\mathbf{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$. We assume that $\widetilde{\boldsymbol{Y}}_{ij}|\mathbf{x}_{ij}$ is independent across ij, and $\widetilde{Y}_{ijk}$ is related to a linear predictor $\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_k$ via

$$\mathrm{pr}\left(\widetilde{Y}_{ijk} = 1|\mathbf{x}_{ij}\right) = g_k\left(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_k\right), \quad \text{for} \quad k = 1, 2, \tag{1}$$

where $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})^{\mathrm{T}}$ is a vector of $(p+1)$ regression coefficients that will be estimated and g_k is a user-chosen known link function (eg, the inverse of the logit or probit link). One could use different links for different infections. Equation (1) builds marginal probability models of the random vector $\widetilde{\boldsymbol{Y}}_{ij}|\mathbf{x}_{ij}$.

In pooled testing, the true infection statuses are often latent due to pooling and potential misclassification. In each group, individual specimens are mixed together to form a pool. We denote the true status of the jth pool by $\widetilde{\boldsymbol{Z}}_j = (\widetilde{Z}_{j1}, \widetilde{Z}_{j2})^{\mathrm{T}}$, where $\widetilde{Z}_{jk} = \max\{\widetilde{Y}_{ijk} : i = 1, \ldots, c_j\}$, ie, $\widetilde{Z}_{jk} = 1$ if the pool involves at least one individual who is positive for the kth infection and $\widetilde{Z}_{jk} = 0$ if otherwise. With the use of an imperfect assay, both $\widetilde{\boldsymbol{Y}}_{ij}$ and $\widetilde{\boldsymbol{Z}}_j$ are latent. Observed data are the testing outcomes from the imperfect multiplex assay. Pools are tested in Stage 1. We denote the testing outcomes of the jth pool by $\boldsymbol{Z}_j = (Z_{j1}, Z_{j2})^{\mathrm{T}}$, where $Z_{jk} = 1(0)$ if the pool tests positive (negative) for the kth infection. If $\boldsymbol{Z}_j = (0, 0)^{\mathrm{T}}$, then

$\boldsymbol{Z}_j$ is the only observed test response for the jth group of individuals. Otherwise, those individuals are tested separately in Stage 2. We denote by $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2})^{\mathrm{T}}$ the retesting outcome of the ith individual in the jth group, ie, $Y_{ijk} = 1(0)$ if the individual is retested as positive (negative) for the kth infection. Note that $\boldsymbol{Y}_{ij}$ can only be observed if $\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}$. In summary, observed testing outcomes from the jth group, denoted by $\boldsymbol{\mathcal{P}}_j$, take one of the two forms, either $\boldsymbol{Z}_j = (0,0)^{\mathrm{T}}$ or $\boldsymbol{Z}_j \in \{(1,0)^{\mathrm{T}}, (0,1)^{\mathrm{T}}, (1,1)^{\mathrm{T}}\}$ and $\boldsymbol{Y}_{1j}, \ldots, \boldsymbol{Y}_{c_j j}$.

The discrepancy between true statuses and testing outcomes is often measured by assay sensitivity and specificity. Denote by $S_{e:k}$ and $S_{p:k}$ the assay sensitivity and specificity, respectively, for the kth infection. In practice, an assay used for large-scale screening is often imperfect. We let $S_{e:k}$ and $S_{p:k}$ be in $(0,1)$. Our methodology posits three assumptions on these misclassification parameters. Assumption 1 is that $S_{e:k}$ and $S_{p:k}$ do not depend on the group size, eg, $S_{e:k} = \mathrm{pr}(Z_{jk} = 1|\widetilde{Z}_{jk} = 1) = \mathrm{pr}(Y_{ijk} = 1|\widetilde{Y}_{ijk} = 1)$ and $S_{p:k} = \mathrm{pr}(Z_{jk} = 0|\widetilde{Z}_{jk} = 0) = \mathrm{pr}(Y_{ijk} = 0|\widetilde{Y}_{ijk} = 0)$ hold for all i, j, and k. Assumption 2 assumes that conditioning on the true statuses of the specimens being tested, testing responses are independent across each other and across infections. Assumption 3 further assumes that given the true statuses, testing responses are independent of the covariates, eg, $\mathrm{pr}(Z_{j1} = 0, Z_{j2} = 1, Y_{ij1} = 1, Y_{ij2} = 0|\widetilde{Z}_{j1} = 0, \widetilde{Z}_{j1} = 0, \widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 1, \boldsymbol{x}_{ij}) = \mathrm{pr}(Z_{j1} = 0|\widetilde{Z}_{j1} = 0)\mathrm{pr}(Z_{j2} = 1|\widetilde{Z}_{j2} = 0)\mathrm{pr}(Y_{ij1} = 1|\widetilde{Y}_{ij1} = 1)\mathrm{pr}(Y_{ij2} = 0|\widetilde{Y}_{ij2} = 1) = S_{p:1}(1 - S_{p:2})S_{e:1}(1 - S_{e:2})$. All these assumptions are standard in group testing literature (see most references in Section 1.2). In practice, one may need to conduct proper assay calibration to ensure the applicability of these assumptions.

Our primary goal is to estimate $\boldsymbol{\beta}_k$, $S_{e:k}$, and $S_{p:k}$. Towards this goal, we want to incorporate the retesting outcomes for two main reasons: (1) Ignoring the retesting outcomes could severely inflate the variance of the estimators of β_k (see the supplementary materials for a numerical illustration), and (2) including the retesting outcomes gives us repeated measurements (ie, many specimens are tested in pools and also individually), which provide valuable information to estimate misclassification parameters. To seamlessly incorporate all retesting outcomes, we propose a *copula-based multivariate binary regression model*. We assume that there exists a vector of standard uniform random variables, $\boldsymbol{U}_{ij} = (U_{ij1}, U_{ij2})^{\mathrm{T}}$, such that the event $\{\widetilde{Y}_{ijk} = 1|\mathbf{x}_{ij}\}$ is equivalent to $\{U_{ijk} \leq g_k(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_k)\}$, where $\boldsymbol{U}_{ij}$ is independent and follows a bivariate copula.[33] Denote the chosen copula by $C\{u_1, u_2|\delta\}$, where $u_1, u_2 \in (0,1)$ and C is known up to a parameter δ (which could be a vector). Then, the marginal regression models in (1) naturally hold, and the coinfection probability is

$$\mathrm{pr}\left(\widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 1|\mathbf{x}_{ij}\right) = C\left\{g_1\left(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1\right), g_2\left(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2\right)|\delta\right\}. \tag{2}$$

Combining (1) and (2) together defines our joint probability model of $\widetilde{\boldsymbol{Y}}_{ij}|\mathbf{x}_{ij}$.

3 | ESTIMATION

We maximize the likelihood function to obtain our estimators of $\boldsymbol{\beta}_k$, $S_{e:k}$, $S_{p:k}$, and δ. For notation simplicity, we write $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \delta)^{\mathrm{T}}$, $\boldsymbol{\theta}_2 = (S_{e:1}, S_{e:2}, S_{p:1}, S_{p:2})^{\mathrm{T}}$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$. Furthermore, we denote by $p_{ijy_1y_2}(\boldsymbol{\theta}_1)$ the cell probability $\mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2|\mathbf{x}_{ij})$ defined by (1) and (2) under $\boldsymbol{\theta}_1$ for $y_1, y_2 \in \{0,1\}$, $i = 1, \ldots, c_j$ and $j = 1, \ldots, J$. Then, $p_{ij11}(\boldsymbol{\theta}_1) = C\{g_1(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1), g_2(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2)|\delta\}$, $p_{ij10}(\boldsymbol{\theta}_1) = g_1(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1) - p_{ij11}(\boldsymbol{\theta}_1)$, $p_{ij01}(\boldsymbol{\theta}_1) = g_2(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2) - p_{ij11}(\boldsymbol{\theta}_1)$, and $p_{ij00}(\boldsymbol{\theta}_1) = 1 - p_{ij11}(\boldsymbol{\theta}_1) - p_{ij10}(\boldsymbol{\theta}_1) - p_{ij01}(\boldsymbol{\theta}_1)$. In the supplementary materials, we derive an expression of the log-likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X})$, where $\boldsymbol{\mathcal{P}}$ and $\mathbf{X}$ denote the collections of $\boldsymbol{\mathcal{P}}_j$ and $\mathbf{x}_{ij}$, respectively. However, due to the complexity of $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X})$, a direct maximization could be time-consuming. The supplementary materials include a numerical illustration of this disadvantage.

We propose a GEM algorithm to accelerate the computation. The algorithm incorporates $\widetilde{\boldsymbol{Y}} = \{\widetilde{\boldsymbol{Y}}_{11}, \ldots, \widetilde{\boldsymbol{Y}}_{c_J J}\}$ as latent variables. The complete log-likelihood function of $\boldsymbol{\theta}$, derived from the conditional distribution of $\boldsymbol{\mathcal{P}}$ and $\widetilde{\boldsymbol{Y}}$ given $\mathbf{X}$, can be written by $\ell_c(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X}) = \ell_{c1}(\boldsymbol{\theta}_1|\widetilde{\boldsymbol{Y}}, \mathbf{X}) + \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}})$, where

$$\begin{aligned}\ell_{c1}(\boldsymbol{\theta}_1|\widetilde{\boldsymbol{Y}}, \mathbf{X}) = \sum_{j=1}^{J}\sum_{i=1}^{c_j}\Big[&(1 - \widetilde{Y}_{ij1})(1 - \widetilde{Y}_{ij2})\log p_{ij00}(\boldsymbol{\theta}_1) + \widetilde{Y}_{ij1}(1 - \widetilde{Y}_{ij2})\log p_{ij10}(\boldsymbol{\theta}_1)\\ &+ (1 - \widetilde{Y}_{ij1})\widetilde{Y}_{ij2}\log p_{ij01}(\boldsymbol{\theta}_1) + \widetilde{Y}_{ij1}\widetilde{Y}_{ij2}\log p_{ij11}(\boldsymbol{\theta}_1)\Big]\end{aligned} \tag{3}$$

and

$$\begin{aligned}
\ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}}) = \sum_{j=1}^{J}\sum_{k=1}^{2}\Bigg[& \left\{\widetilde{Z}_{jk}Z_{jk} + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}\widetilde{Y}_{ijk}Y_{ijk}\right\}\log S_{e:k} \\
& + \left\{\widetilde{Z}_{jk}(1-Z_{jk}) + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}\widetilde{Y}_{ijk}(1-Y_{ijk})\right\}\log(1-S_{e:k}) \\
& + \left\{(1-\widetilde{Z}_{jk})(1-Z_{jk}) + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}(1-\widetilde{Y}_{ijk})(1-Y_{ijk})\right\}\log S_{p:k} \\
& + \left\{(1-\widetilde{Z}_{jk})Z_{jk} + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}(1-\widetilde{Y}_{ijk})Y_{ijk}\right\}\log(1-S_{p:k})\Bigg], \qquad (4)
\end{aligned}$$

in which $\widetilde{Z}_{jk} = \max\{\widetilde{Y}_{ijk} : i = 1, \ldots, c_j\}$ and $I(\cdot)$ is the indicator function.

Our GEM algorithm starts at an initial value and then iterates between an E-step and an M-step to update the value until reaching a numerical convergence. At a current value $\boldsymbol{\theta}^{(d)}$, the E-step calculates $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(d)}) = \mathcal{Q}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^{(d)}) + \mathcal{Q}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^{(d)})$, where $\mathcal{Q}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^{(d)}) = E\{\ell_{c1}(\boldsymbol{\theta}_1|\widetilde{\mathbf{Y}},\mathbf{X})|\boldsymbol{\mathcal{P}},\mathbf{X},\boldsymbol{\theta}^{(d)}\}$ and $\mathcal{Q}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^{(d)}) = E\{\ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\mathbf{Y}})|\boldsymbol{\mathcal{P}},\mathbf{X},\boldsymbol{\theta}^{(d)}\}$. After an inspection of (3) and (4), it suffices to calculate $\eta_{ij00}^{(d)}, \eta_{ij10}^{(d)}, \eta_{ij01}^{(d)}, \eta_{ij11}^{(d)}$ (for $\mathcal{Q}_1$), and $\eta_{\boldsymbol{\mathcal{P}},jk}^{(d)}$ (for $\mathcal{Q}_2$), where

$$\eta_{ijy_1y_2}^{(d)} = \mathrm{pr}\left(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2|\boldsymbol{\mathcal{P}},\mathbf{X},\boldsymbol{\theta}^{(d)}\right) \text{ and } \eta_{\boldsymbol{\mathcal{P}},jk}^{(d)} = \mathrm{pr}\left(\widetilde{Z}_{jk} = 1|\boldsymbol{\mathcal{P}},\mathbf{X},\boldsymbol{\theta}^{(d)}\right), \qquad (5)$$

for $i = 1, \ldots, c_j, j = 1, \ldots, J, y_1, y_2 \in \{0, 1\}$, and $k = 1, 2$. Although $\eta_{ijy_1y_2}^{(d)}$ values have been studied without the consideration of $\mathbf{X}$[26], they were not updated in closed forms; thus, a Gibbs sampler was employed to approximate these quantities. However, in the regression context, using such approximations requires enlarging the tolerance of the numerical convergence and hence might induce bias. To improve the computational accuracy, we calculate all the probabilities in (5) exactly (see the supplementary materials for details).

With the probabilities in (5) calculated, we rewrite $\mathcal{Q}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^{(d)})$ by

$$\mathcal{Q}_1\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\delta|\boldsymbol{\theta}^{(d)}\right) = \sum_{j=1}^{J}\sum_{i=1}^{c_j}\sum_{y_1=0}^{1}\sum_{y_2=0}^{1}\eta_{ijy_1y_2}^{(d)}\log p_{ijy_1y_2}(\boldsymbol{\theta}_1),$$

and $\mathcal{Q}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^{(d)})$ by

$$\sum_{k=1}^{2}\left\{W_{1k}^{(d)}\log S_{e:k} + W_{2k}^{(d)}\log(1-S_{e:k}) + W_{3k}^{(d)}\log S_{p:k} + W_{4k}^{(d)}\log(1-S_{p:k})\right\}, \qquad (6)$$

where

$$\begin{aligned}
W_{1k}^{(d)} &= \sum_{j=1}^{J}\left\{\eta_{\boldsymbol{\mathcal{P}},jk}^{(d)}Z_{jk} + I\left(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}\right)\sum_{i=1}^{c_j}\eta_{ij,k}^{(d)}Y_{ijk}\right\}, \\
W_{2k}^{(d)} &= \sum_{j=1}^{J}\left\{\eta_{\boldsymbol{\mathcal{P}},jk}^{(d)}(1-Z_{jk}) + I\left(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}\right)\sum_{i=1}^{c_j}\eta_{ij,k}^{(d)}(1-Y_{ijk})\right\}, \\
W_{3k}^{(d)} &= \sum_{j=1}^{J}\left\{\left(1-\eta_{\boldsymbol{\mathcal{P}},jk}^{(d)}\right)(1-Z_{jk}) + I\left(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}\right)\sum_{i=1}^{c_j}\left(1-\eta_{ij,k}^{(d)}\right)(1-Y_{ijk})\right\}, \\
W_{4k}^{(d)} &= \sum_{j=1}^{J}\left\{\left(1-\eta_{\boldsymbol{\mathcal{P}},jk}^{(d)}\right)Z_{jk} + I\left(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}\right)\sum_{i=1}^{c_j}\left(1-\eta_{ij,k}^{(d)}\right)Y_{ijk}\right\},
\end{aligned}$$

in which, $\eta_{ij,1}^{(d)} = \eta_{ij11}^{(d)} + \eta_{ij10}^{(d)}$ and $\eta_{ij,2}^{(d)} = \eta_{ij11}^{(d)} + \eta_{ij01}^{(d)}$. The M-step in our GEM algorithm updates $\boldsymbol{\theta}_1^{(d)}$ by $\boldsymbol{\theta}_1^{(d+1)} = (\boldsymbol{\beta}_1^{(d+1)^{\mathrm{T}}}(\boldsymbol{\beta}_2^{(d+1)^{\mathrm{T}}}\delta^{(d+1)})^{\mathrm{T}}$ where $\boldsymbol{\beta}_1^{(d+1)} = \operatorname{argmax}_{\boldsymbol{\beta}_1}\ \mathcal{Q}_1(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^{(d)}, \delta^{(d)}|\boldsymbol{\theta}^{(d)})$, $\boldsymbol{\beta}_2^{(d+1)} = \operatorname{argmax}_{\boldsymbol{\beta}_2}\ \mathcal{Q}_1(\boldsymbol{\beta}_1^{(d+1)}, \boldsymbol{\beta}_2, \delta^{(d)}|\boldsymbol{\theta}^{(d)})$, and $\delta^{(d+1)} = \operatorname{argmax}_\delta \mathcal{Q}_1(\boldsymbol{\beta}_1^{(d+1)}, \boldsymbol{\beta}_2^{(d+1)}, \delta|\boldsymbol{\theta}^{(d)})$. The value of $\boldsymbol{\theta}_2^{(d+1)}$ is obtained by maximizing (6) and can be written as

$$\boldsymbol{\theta}_2^{(d+1)} = \left(S_{e:1}^{(d+1)}, S_{e:2}^{(d+1)}, S_{p:1}^{(d+1)}, S_{p:2}^{(d+1)}\right)^{\mathrm{T}},$$

where $S_{e:k}^{(d+1)} = W_{1k}^{(d)}/(W_{1k}^{(d)} + W_{2k}^{(d)})$ and $S_{p:k}^{(d+1)} = W_{3k}^{(d)}/(W_{3k}^{(d)} + W_{4k}^{(d)})$, for $k = 1, 2$. Combining $\boldsymbol{\theta}_1^{(d+1)}$ and $\boldsymbol{\theta}_2^{(d+1)}$ provides $\boldsymbol{\theta}^{(d+1)}$. Because $\mathcal{Q}(\boldsymbol{\theta}^{(d+1)}|\boldsymbol{\theta}^{(d)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(d)}|\boldsymbol{\theta}^{(d)})$, the convergence of $\{\boldsymbol{\theta}^{(d)}\}_{d=1}^{\infty}$ is guaranteed.[30] We denote by $\hat{\boldsymbol{\theta}}$ the limit of $\boldsymbol{\theta}^{(d)}$.

Denote by $\mathcal{I}(\boldsymbol{\theta})$ the observed data information matrix. Following the standard arguments of the MLE,[34] we have $\mathcal{I}(\widehat{\boldsymbol{\theta}})^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to $\mathcal{N}(0, \boldsymbol{I}_{2p+7})$ as $N \to \infty$, where $\boldsymbol{I}_m$ denotes the m-dimensional identity matrix. Applying Louis' method[35] provides

$$\mathcal{I}(\boldsymbol{\theta}) = -E\left\{ \frac{\partial^2 \ell_c(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \widetilde{\mathbf{Y}}, \mathbf{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} \middle| \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta} \right\} - \operatorname{cov}\left\{ \frac{\partial \ell_c(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \widetilde{\mathbf{Y}}, \mathbf{X})}{\partial\boldsymbol{\theta}} \middle| \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta} \right\}.$$

Again, instead of approximating $\mathcal{I}(\boldsymbol{\theta})$ via the Gibbs sampling approach,[26] we are able to calculate it exactly. The calculations are included in the supplementary materials. With $\mathcal{I}(\widehat{\boldsymbol{\theta}})$, one can make large sample Wald-type inferences. For example, let θ_l, $\widehat{\theta}_l$, and $\widehat{\sigma}_{ll}^2$ be the lth component of $\boldsymbol{\theta}$, the lth component of $\widehat{\boldsymbol{\theta}}$, and the lth diagonal entry of $\mathcal{I}(\widehat{\boldsymbol{\theta}})^{-1}$, respectively, for $l = 1, \dots, 2p + 7$. The estimated standard error (SE) of $\widehat{\theta}_l$ is $\widehat{\sigma}_{ll}$ and an approximated $100(1 - \alpha)\%$ confidence interval of θ_l is $\widehat{\theta}_l \pm z_{\alpha/2}\widehat{\sigma}_{ll}$, where z_α is the αth upper quantile of $\mathcal{N}(0, 1)$.

4 | VARIABLE SELECTION FOR EACH INFECTION

With $\widehat{\boldsymbol{\theta}}$ and $\mathcal{I}(\widehat{\boldsymbol{\theta}})$ computed, we further identify which risk factors are truly relevant for each infection. Denote by $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ the values of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ that generate the true individual statuses $\widetilde{\boldsymbol{Y}}$, respectively, where $\boldsymbol{\beta}_k^* = (\beta_{k0}^*, \beta_{k1}^*, \dots, \beta_{kp}^*)^{\mathrm{T}}$ for $k = 1, 2$. One can index the significant risk factors to the kth infection by $\mathcal{M}_k = \{j \in \mathcal{M} : \beta_{kj}^* \neq 0\}$, where we take $\mathcal{M} = \{1, 2, \dots, p\}$ by defaulting that an intercept term is always included in the model. One must note that $\mathcal{M}_1$ and $\mathcal{M}_2$ might be different.

We apply a shrinkage method to simultaneously select $\mathcal{M}_k$ and estimate nonzero β_{kj}^*. To unify notation, we write $\boldsymbol{\theta}_{\mathcal{T}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}\mathcal{T}}$ as the subvector and the submatrix of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\Sigma}}$ according to an index set $\mathcal{T} \subset \{1, \dots, 2p + 7\}$, respectively. Let $\mathcal{A} = \{2, \dots, p + 1, p + 3, \dots, 2p + 2\}$. Our shrinkage estimator of $\boldsymbol{\theta}_{\mathcal{A}}$ is defined by

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda} = \underset{\boldsymbol{\theta}_{\mathcal{A}}}{\operatorname{argmin}} \left\{ \frac{1}{2}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{\mathcal{A}})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{\mathcal{A}}) + \sum_{k=1}^{2}\lambda_k\sum_{j=1}^{p}\omega_{kj}|\beta_{kj}| \right\}, \tag{7}$$

where $\lambda_k \sum_{j=1}^{p} \omega_{kj}|\beta_{kj}|$ is an adaptive LASSO penalty,[36] $\lambda_k \geq 0$ is a tuning parameter that controls the shrinkage level, and $\omega_{kj} = |\widehat{\beta}_{kj}|^{-1}$ is an adaptive weight. When λ_k is 0, $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda} = \widehat{\boldsymbol{\theta}}_{\mathcal{A}}$. When λ_k increase, due to the singularity of the absolute value function at the origin, components of $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}$ are penalized to zero one by one. Writing $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda} = (\widetilde{\beta}_{11,\lambda}, \dots, \widetilde{\beta}_{1p,\lambda}, \widetilde{\beta}_{21,\lambda}, \dots, \widetilde{\beta}_{2p,\lambda})^{\mathrm{T}}$, we estimate $\mathcal{M}_1$ and $\mathcal{M}_2$ by $\widetilde{\mathcal{M}}_{1,\lambda} = \{j \in \mathcal{M} : \widetilde{\beta}_{1j,\lambda} \neq 0\}$ and $\widetilde{\mathcal{M}}_{2,\lambda} = \{j \in \mathcal{M} : \widetilde{\beta}_{2j,\lambda} \neq 0\}$, respectively.

Computing $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}$ is fast. The objective function in (7) is simply a summation of a quadratic function and a weighted l_1-norm of $\boldsymbol{\theta}_{\mathcal{A}}$ and therefore can be quickly minimized by slightly modifying the seminal least angle regression.[37] Let $\mathcal{A}^c = \{1, 2, \dots, 2p + 7\} \setminus \mathcal{A}$ and $\ell(\boldsymbol{\theta}_{\mathcal{A}}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \widehat{\boldsymbol{\theta}}_{\mathcal{A}^c})$ be the log-likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X})$ with $\boldsymbol{\theta}_{\mathcal{A}^c}$ fixed to be $\widehat{\boldsymbol{\theta}}_{\mathcal{A}^c}$. One could also construct a shrinkage estimator by the traditional penalized MLE,[38] which minimizes $-\ell(\boldsymbol{\theta}_{\mathcal{A}}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \widehat{\boldsymbol{\theta}}_{\mathcal{A}^c}) + \sum_{k=1}^{2} \lambda_k \sum_{j=1}^{p} \omega_{kj}|\beta_{kj}|$. As the quadratic term in (7) is the leading component of the Taylor's expansion of $-\ell(\boldsymbol{\theta}_{\mathcal{A}}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \widehat{\boldsymbol{\theta}}_{\mathcal{A}^c})$ at $\boldsymbol{\theta}_{\mathcal{A}} = \widehat{\boldsymbol{\theta}}_{\mathcal{A}}$, it can be easily shown that $\widetilde{\boldsymbol{\theta}}_{\mathcal{A}}$ and the penalized MLE are asymptotically equivalent. However, the computation cost of obtaining penalized MLE will be a lot higher due to the complexity of the log-likelihood function.

The use of adaptive weights ω_{kj} is critical to achieve the *oracle* properties.[36] It assigns sufficiently large penalties to insignificant covariates so that they would be excluded from the model; on the other hand, it imposes mild penalties to significant ones in order that they would be retained in the model. The oracle properties are stated as follows. As $N \to \infty$, if $\max(\lambda_1, \lambda_2)/\sqrt{N} \to 0$ and $\min(\lambda_1, \lambda_2) \to \infty$, we have both the selection consistency $\text{pr}(\widetilde{\mathcal{M}}_{1,\lambda} = \mathcal{M}_1, \widetilde{\mathcal{M}}_{2,\lambda} = \mathcal{M}_2) \to 1$ and the estimation consistency $\sup_{k,j} \|\widetilde{\beta}_{kj,\lambda} - \beta^*_{kj}\| = O_p(N^{-1/2})$. The proof follows similar arguments in the proofs of Theorems 1 and 2 in Wang and Leng[39] and thus is omitted.

To select λ_1 and λ_2, we propose to minimize a type of Bayesian information criterion (BIC),[40] ie,

$$\text{BIC}(\lambda_1, \lambda_2) = (\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda})^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} (\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}) + \{df_{1,\lambda} + df_{2,\lambda}\} \log N, \tag{8}$$

where $df_{k,\lambda} = |\widetilde{\mathcal{M}}_{k,\lambda}|$ for $k = 1, 2$. Following the proof of theorem 3 in the work of Wang et al,[41] one can show that, with the optimal (λ_1, λ_2) from (8), $\text{pr}(\widetilde{\mathcal{M}}_{1,\lambda} = \mathcal{M}_1, \widetilde{\mathcal{M}}_{2,\lambda} = \mathcal{M}_2) \to 1$ as $N \to \infty$. In other words, any (λ_1, λ_2) that does not lead to the correct variable selection cannot be selected by (8) when the number of individuals is large.

The purpose of this subsection is to provide a shrinkage estimator of the regression coefficients, of which the sparsity pattern can help us identify the truly relevant risk factor for each infection. Inference procedures, such as constructing a confidence interval or conducting hypothesis testing, based on this shrinkage estimator are beyond the scope of this work. There are numerous studies demonstrating that even in classical linear regression, finite-sample inference procedures based on asymptotic properties of the adaptive LASSO estimator perform poorly.[42] Developing valid inferential methods for shrinkage estimators in group testing, even with a single infection, could be an interesting but challenging future research topic. In this article, it is the variable selection of primary interest.

5 | NUMERICAL STUDIES

5.1 | Simulation

We consider three different settings for the joint distribution of $\widetilde{\boldsymbol{Y}}_{ij}|\mathbf{x}_{ij}$. In all of them, we keep both g_1 and g_2 in the marginal regression model (1) being the inverse of the logit link function, and use a Gumbel copula,[43] as shown in the following: $C(u_1, u_2|\delta) = \exp\{-[(-\log u_1)^{1/\delta} + (-\log u_2)^{1/\delta}]^{\delta}\}$ with $\delta = 0.3$, to generate the coinfection probability (2). The difference across the three settings comes from the choices of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{x})$, where $\mathbf{x}$ is a generic notation of $\mathbf{x}_{ij}$:

- (S1) $\boldsymbol{\beta}_1 = (-5, -3, 2, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (-5, -3, 0, 3, 0, 0)^{\mathrm{T}}$, and $\mathbf{x} = (1, x_1, \dots, x_5)^{\mathrm{T}}$, where we independently simulate x_1 from $\mathcal{N}(0, 1)$, x_2, and x_3 from Bernoulli(0.4), x_4 from Uniform$(-0.5, 0.5)$, and x_5 from $\mathcal{N}(0, 0.75^2)$.
- (S2) $\boldsymbol{\beta}_1 = (-4, -2, 2, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (-5, -2, 0, -2, 0, 0)^{\mathrm{T}}$, and $\mathbf{x} = (1, x_1, \dots, x_5)^{\mathrm{T}}$, where $\mathbf{x}$ is simulated from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ with $[\boldsymbol{\Omega}]_{st} = 1$ if $s = t$ and $[\boldsymbol{\Omega}]_{st} = 0.5$ if $s \neq t$.
- (S3) $\boldsymbol{\beta}_1 = (-5, (-2, -2, -2, 2, 2) \otimes (1, 0))^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (-6, (-3, -3, 2, 3, 0) \otimes (1, 0))^{\mathrm{T}}$, and $\mathbf{x} = (1, x_1, \dots, x_{10})^{\mathrm{T}}$, where $\otimes$ is the Kronecker product, $\mathbf{x}$ is simulated from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ with $[\boldsymbol{\Omega}]_{st} = 1$ if $s = t$ and $[\boldsymbol{\Omega}]_{st} = 0.5$ if $s \neq t$.

Note that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have different sparsity patterns (eg, in S1, x_2 is significant to the first infection but not to the second infection). This is to emulate the situation where two infections have different sets of significant risk factors. The values of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are chosen in a way such that the prevalence of each infection is about 7%-10%.

Under each setting, we simulate two types of data: individual testing data and the SHL-pooled testing data. To do so, we first generate $N = 3000$ individual covariates. Given a set of covariates, we calculate the individual's cell probabilities $(p_{ijy_1y_2})$ using the specified copula-based multivariate binary regression model and then generate the true infection statuses for both infections from a multinomial distribution with those cell probabilities. We denote the covariates and the true infection statuses of the nth individual by $\mathbf{x}_n$ and $\widetilde{\boldsymbol{Y}}_n = (\widetilde{Y}_{n1}, \widetilde{Y}_{n2})^{\mathrm{T}}$, respectively, for $n = 1, \dots, 3000$. Herein, because groups have not been created yet, we use the subscript n instead of the ij (in $\widetilde{\boldsymbol{Y}}_{ij}$ and $\mathbf{x}_{ij}$). Given $(\widetilde{\boldsymbol{Y}}_n, \mathbf{x}_n)$, we simulate individual testing data and the SHL-pooled testing data. We let $S_{e:k} = S_{p:k} = 0.95$ for $k = 1, 2$. Values other than 0.95 are considered in the supplementary materials.

Based on $\widetilde{\boldsymbol{Y}}_n$, we generate individual testing outcomes of the nth specimen by $\boldsymbol{T}_n = (T_{n1}, T_{n2})^{\mathrm{T}}$, where $T_{nk} \sim$ Bernoulli$\{S_{e:k}\widetilde{Y}_{nk} + (1 - S_{p:k})(1 - \widetilde{Y}_{nk})\}$. Then, we estimate $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta)^{\mathrm{T}}$ from $(\boldsymbol{T}_n, \mathbf{x}_n)$. This estimation procedure is similar to the one outlined in Section 3. We also use a GEM algorithm to compute the MLEs and Louis' method to calculate the observed data information matrix for making large-sample Wald-type inferences. Furthermore, we slightly modify our variable selection method (in Section 4) to accommodate individual testing data. All the details are provided in the supplementary materials. It is worthwhile to note that $S_{e:k}$ and $S_{p:k}$ could not be estimated in individual testing data. Hence, with individual testing data $(\boldsymbol{T}_n, \mathbf{x}_n)$, we have to assume the true values of $S_{e:k}$ and $S_{p:k}$ as known to estimate $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta)$.

We generate the SHL-pooled testing data from $\widetilde{\boldsymbol{Y}}_n$. A common group size is used in our simulations, ie, $c_j = c$ and $c \in \{2, 5, 10\}$. For a fixed c, we randomly assign the 3000 individuals to one of $J = 3000/c$ groups. With the group membership identified, we relabel $(\widetilde{\boldsymbol{Y}}_n, \mathbf{x}_n)$ by $(\widetilde{\boldsymbol{Y}}_{ij}, \mathbf{x}_{ij})$, where $i = 1, \dots, c$ and $j = 1, \dots, J$. The true statuses of the jth pool are calculated as $\widetilde{Z}_{jk} = \max_i \widetilde{Y}_{ijk}$ where $k = 1, 2$. Then, we generate the pooled testing outcomes by $\boldsymbol{Z}_j = (Z_{j1}, Z_{j2})^{\mathrm{T}}$, where $Z_{jk} \sim \text{Bernoulli}\{S_{e:k}\widetilde{Z}_{jk} + (1 - S_{p:k})(1 - \widetilde{Z}_{jk})\}$. As per the SHL pooling protocol, only if $\max(Z_{j1}, Z_{j2}) = 1$, we generate retesting outcomes of the ith individual in this group by $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2})^{\mathrm{T}}$, where $Y_{ijk} \sim \text{Bernoulli}\{S_{e:k}\widetilde{Y}_{ijk} + (1 - S_{p:k})(1 - \widetilde{Y}_{ijk})\}$. Collecting all $\boldsymbol{Z}_j$ and $\boldsymbol{Y}_{ij}$ yields the SHL-pooled testing data $\boldsymbol{\mathcal{P}}$. Note that the number of tests that were used to obtain $\boldsymbol{\mathcal{P}}$ is the summation of J and the number of $\boldsymbol{Y}_{ij}$. From $\boldsymbol{\mathcal{P}}$ and $\mathbf{x}_{ij}$, we estimate $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta, S_{e:1}, S_{e:2}, S_{p:1}, S_{p:2})$.

We repeat 500 times the process of generating $\boldsymbol{T}_n$ and $\boldsymbol{\mathcal{P}}$ for each $c \in \{2, 5, 10\}$. For each set of individual testing data or the SHL-pooled testing data, we first treat the diagnosis results for each infection as the true statues and fit them using our copula-based multivariate binary regression model. The resulting MLE of $\boldsymbol{\theta}_1$ is used as the initial value of $\boldsymbol{\theta}_1$. The initial values of the assay sensitivity and specificity are chosen to be 0.9. Then, we run our GEM algorithm to compute the MLE and use Louis' method to construct a 95% confidence interval for each unknown parameter (see the last paragraph of Section 3). In addition to the BIC-type shrinkage estimator, we also compute an Akaike information criterion (AIC)–type[44] and an extended regularized information criterion (ERIC)–type[45] estimator using the tuning parameters selected by minimizing $\text{AIC}(\lambda_1, \lambda_2) = (\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}) + 2\{df_{1,\lambda} + df_{2,\lambda}\}$ and $\text{ERIC}(\lambda_1, \lambda_2) = (\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}) + df_{1,\lambda}\log(N/\lambda_1) + df_{2,\lambda}\log(N/\lambda_2)$, Respectively. For individual testing data, slightly modified versions are available in the supplementary materials.

To compare the overall performance of the MLE and three shrinkage estimators, we consider the prediction error $\text{PE} = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\{\sum_{y_1=0}^{1}\sum_{y_2=0}^{1}(\hat{p}_{ijy_1y_2} - p^*_{ijy_1y_2})^2\}^{1/2}$, where $p^*_{ijy_1y_2}$ are the true cell probabilities and $\hat{p}_{ijy_1y_2}$ are the predicted cell probabilities using an estimator of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta)$. To evaluate the variable selection performance of shrinkage estimators, we define by the selection rate (SR) the proportion of the true model being exactly selected by a shrinkage estimator. Results from the 500 replications under S1–S3 are summarized in Tables 1–4.

Tables 1 to 3 provide summary statistics of the MLEs for S1-S3, respectively. Under both individual testing and the SHL pooling protocol, the MLEs of the unknown parameters obtained by our GEM algorithm exhibit little, if any, evidence of

TABLE 1 Summary statistics of the 500 maximum likelihood estimates obtained under S1, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE), and the empirical coverage (EC) of 95% confidence intervals under either individual testing (IT) or the State Hygienic Laboratory pooling with $c = 2, 5, 10$. The average number of tests (# of tests) under each protocol is also provided. The prevalence (averaged over 500 repetitions) of the first and second infections are 7.64% and 8.22%, respectively

		IT		**$c = 2$**		**$c = 5$**		**$c = 10$**	
# tests		**3000**		**2351**		**2078**		**2445**	
	Truth	**Mean(SD)**	**EC(SE)**	**Mean(SD)**	**EC(SE)**	**Mean(SD)**	**EC(SE)**	**Mean(SD)**	**EC(SE)**
β_{10}	-5	-5.08(0.36)	0.94(0.37)	-5.06(0.29)	0.94(0.29)	-5.06(0.31)	0.94(0.29)	-5.07(0.34)	0.95(0.32)
β_{11}	-3	-3.05(0.25)	0.96(0.26)	-3.03(0.21)	0.94(0.21)	-3.04(0.22)	0.94(0.21)	-3.04(0.24)	0.95(0.23)
β_{12}	2	2.03(0.27)	0.94(0.27)	2.02(0.24)	0.94(0.24)	2.02(0.25)	0.94(0.24)	2.03(0.26)	0.95(0.25)
β_{13}	0	-0.01(0.24)	0.95(0.23)	-0.01(0.22)	0.95(0.21)	-0.01(0.22)	0.95(0.21)	-0.01(0.22)	0.94(0.21)
β_{14}	0	0.01(0.38)	0.95(0.39)	0.01(0.34)	0.96(0.35)	-0.01(0.35)	0.96(0.35)	0.00(0.37)	0.96(0.36)
β_{15}	0	0.00(0.19)	0.96(0.20)	0.00(0.17)	0.97(0.18)	0.00(0.17)	0.97(0.18)	0.00(0.19)	0.94(0.19)
β_{20}	-5	-5.08(0.37)	0.95(0.37)	-5.05(0.28)	0.94(0.30)	-5.05(0.30)	0.94(0.30)	-5.04(0.33)	0.96(0.32)
β_{21}	-3	-3.04(0.26)	0.95(0.26)	-3.03(0.21)	0.94(0.22)	-3.03(0.22)	0.94(0.21)	-3.02(0.24)	0.95(0.23)
β_{22}	0	-0.01(0.24)	0.94(0.23)	-0.01(0.21)	0.93(0.21)	0.00(0.22)	0.93(0.21)	-0.01(0.23)	0.94(0.21)
β_{23}	3	3.04(0.33)	0.94(0.32)	3.03(0.27)	0.94(0.27)	3.03(0.29)	0.94(0.27)	3.03(0.30)	0.94(0.29)
β_{24}	0	0.00(0.40)	0.95(0.38)	0.02(0.34)	0.95(0.35)	0.01(0.35)	0.95(0.35)	0.00(0.36)	0.96(0.36)
β_{25}	0	0.01(0.20)	0.95(0.20)	0.00(0.18)	0.94(0.18)	0.00(0.18)	0.94(0.18)	0.00(0.19)	0.95(0.19)
δ	0.3	0.28(0.09)	0.97(0.10)	0.29(0.06)	0.95(0.06)	0.29(0.06)	0.95(0.06)	0.29(0.07)	0.95(0.07)
$S_{e:1}$	0.95	–	–	0.95(0.02)	0.93(0.02)	0.95(0.02)	0.93(0.02)	0.95(0.02)	0.90(0.02)
$S_{e:2}$	0.95	–	–	0.95(0.01)	0.95(0.01)	0.95(0.02)	0.91(0.01)	0.95(0.02)	0.92(0.02)
$S_{p:1}$	0.95	–	–	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)
$S_{p:2}$	0.95	–	–	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.93(0.01)

TABLE 2 Summary statistics of the 500 maximum likelihood estimates obtained under S2, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE), and the empirical coverage (EC) of 95% confidence intervals under either individual testing (IT) or the State Hygienic Laboratory pooling with $c = 2, 5, 10$. The average number of tests (# of tests) under each protocol is also provided. The prevalence (averaged over 500 repetitions) of the first and the second infections are 6.77% and 9.98%, respectively

		IT		$c = 2$		$c = 5$		$c = 10$	
# tests		3000		2493		2312		2678	
	Truth	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)
β_{10}	-4	-4.05(0.25)	0.95(0.26)	-4.02(0.20)	0.95(0.19)	-4.03(0.19)	0.97(0.20)	-4.03(0.21)	0.96(0.21)
β_{11}	-2	-2.03(0.19)	0.96(0.20)	-2.02(0.15)	0.96(0.15)	-2.03(0.16)	0.96(0.16)	-2.02(0.17)	0.95(0.17)
β_{12}	2	2.03(0.19)	0.94(0.20)	2.02(0.16)	0.95(0.16)	2.02(0.16)	0.96(0.16)	2.02(0.17)	0.95(0.17)
β_{13}	0	0.00(0.13)	0.95(0.14)	0.00(0.12)	0.95(0.12)	0.00(0.12)	0.96(0.12)	0.00(0.12)	0.95(0.13)
β_{14}	0	0.00(0.13)	0.96(0.14)	0.00(0.12)	0.95(0.12)	0.00(0.12)	0.96(0.12)	-0.01(0.13)	0.95(0.13)
β_{15}	0	0.00(0.14)	0.95(0.14)	0.00(0.11)	0.96(0.12)	0.00(0.12)	0.95(0.12)	0.00(0.13)	0.95(0.13)
β_{20}	-5	-5.06(0.36)	0.94(0.35)	-5.04(0.26)	0.96(0.27)	-5.03(0.28)	0.97(0.29)	-5.04(0.32)	0.94(0.33)
β_{21}	-2	-2.04(0.20)	0.95(0.20)	-2.03(0.16)	0.97(0.17)	-2.02(0.17)	0.97(0.17)	-2.03(0.19)	0.95(0.19)
β_{22}	0	0.01(0.13)	0.97(0.13)	0.00(0.12)	0.95(0.12)	0.01(0.12)	0.96(0.12)	0.01(0.13)	0.95(0.13)
β_{23}	-2	-2.04(0.20)	0.95(0.20)	-2.03(0.17)	0.96(0.17)	-2.02(0.17)	0.95(0.17)	-2.02(0.19)	0.94(0.18)
β_{24}	0	0.01(0.14)	0.93(0.13)	0.01(0.12)	0.93(0.12)	0.00(0.12)	0.94(0.12)	0.00(0.13)	0.95(0.13)
β_{25}	0	0.01(0.13)	0.95(0.13)	0.01(0.12)	0.95(0.12)	0.00(0.12)	0.95(0.12)	0.01(0.12)	0.96(0.13)
δ	0.3	0.30(0.08)	0.99(0.11)	0.30(0.06)	0.97(0.07)	0.30(0.07)	0.97(0.07)	0.30(0.08)	0.97(0.08)
$S_{e:1}$	0.95	–	–	0.95(0.02)	0.93(0.02)	0.95(0.02)	0.93(0.02)	0.95(0.02)	0.92(0.02)
$S_{e:2}$	0.95	–	–	0.95(0.02)	0.95(0.01)	0.95(0.02)	0.92(0.01)	0.95(0.02)	0.91(0.02)
$S_{p:1}$	0.95	–	–	0.95(0.01)	0.97(0.01)	0.95(0.01)	0.95(0.01)	0.95(0.01)	0.92(0.01)
$S_{p:2}$	0.95	–	–	0.95(0.01)	0.92(0.01)	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.92(0.01)

bias, across all considered settings. Regarding the use of Louis' method, we notice that the average SEs are in agreement with the sample standard deviations of the estimates. In addition, the empirical coverage probabilities for 95% confidence intervals are predominantly at the nominal level. These results indicate that the observed data information matrix is estimated correctly via Louis' method.

To examine the performance of the variable selection, Table 4 provides the SR (in parenthesis) of each shrinkage estimator across all considered settings. One can see that our BIC-type estimator performs the best in identifying the true model in each scenario. For example, in S3 when $c = 2$, the SR using the BIC criterion is 0.820, which is significantly larger than the ones using the AIC (0.294) and the ERIC (0.448). These results demonstrate the advantage of using the BIC in identifying risk factors that are truly relevant for each infection.

Table 4 also provides the average PE×100 values of the MLE and the three shrinkage estimators across all settings. It is clear that all the shrinkage estimators produce smaller prediction errors than the MLE. For example, the BIC-type estimator can reduce almost 50% of the prediction error of the MLE. This is because the adaptive LASSO penalty in (7) could eliminate unnecessary risk factors. Furthermore, because our BIC-type estimator outperforms the other two in term of variable selection, its prediction errors are the smallest under all settings. In conclusion, using the BIC-type shrinkage estimator not only provides a large chance of identifying truly relevant covariates but also yields a high prediction accuracy.

Finally, we want to see whether the SHL pooling protocol causes a loss of information and thus compromises regression inference, when compared to individual testing. To find the answer, we revisit Tables 1–4. This time, we focus on the comparison between individual testing and the SHL pooling. Tables 1 to 3 provide the average number of tests under each setting. Obviously, the SHL pooling protocol uses fewer tests than individual testing (saves about 16% costs). This is an expected appealing feature of the SHL pooling.[26] Also, we observe more: (1) In Tables 1 to 3, the standard deviations obtained using pooling data are uniformly less than the ones obtained using individual testing, suggesting that the SHL pooling could provide a less variational MLE; (2) all the averaged SEs under the SHL pooling are smaller than the ones under individual testing, meaning that one could use the SHL-pooled testing data to construct narrower confidence intervals while maintaining the same nominal level; (3) the advantage of pooling also holds when comparing the average PE×100 values in Table 4, indicating that the SHL pooling enables one to make a better prediction of an individual's infection probabilities; and (4) in terms of variable selection, the highest SR value (in Table 4) always occurs at $c > 1$ under each setting, that is, using the SHL-pooled testing data has a larger chance to identify the true model. Hence, instead of

TABLE 3 Summary statistics of the 500 maximum likelihood estimates obtained under S3, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE), and the empirical coverage (EC) of 95% confidence intervals under either individual testing (IT) or the State Hygienic Laboratory pooling with $c = 2, 5, 10$. The average number of tests (# of tests) under each protocol is also provided. The prevalence (averaged over 500 repetitions) of the first and the second infections are 9.97% and 8.54%, respectively

		IT		$c = 2$		$c = 5$		$c = 10$	
# tests		3000		2508		2337		2701	
	Truth	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)
β_{10}	-5	-5.10(0.39)	0.94(0.36)	-5.07(0.30)	0.93(0.27)	-5.10(0.33)	0.92(0.30)	-5.09(0.36)	0.95(0.34)
β_{11}	-2	-2.04(0.22)	0.94(0.21)	-2.03(0.18)	0.95(0.17)	-2.05(0.20)	0.93(0.18)	-2.04(0.20)	0.94(0.20)
β_{12}	0	0.00(0.13)	0.97(0.14)	0.00(0.12)	0.98(0.13)	0.00(0.13)	0.96(0.13)	0.00(0.13)	0.97(0.14)
β_{13}	-2	-2.04(0.22)	0.93(0.21)	-2.03(0.18)	0.94(0.17)	-2.04(0.19)	0.94(0.18)	-2.04(0.21)	0.93(0.20)
β_{14}	0	0.01(0.14)	0.96(0.14)	0.00(0.12)	0.95(0.13)	0.00(0.13)	0.94(0.13)	0.00(0.13)	0.97(0.14)
β_{15}	-2	-2.05(0.22)	0.94(0.21)	-2.04(0.18)	0.94(0.17)	-2.05(0.19)	0.93(0.18)	-2.05(0.21)	0.92(0.19)
β_{16}	0	0.01(0.14)	0.96(0.14)	0.00(0.13)	0.95(0.13)	0.01(0.13)	0.96(0.13)	0.01(0.14)	0.95(0.14)
β_{17}	2	2.04(0.22)	0.95(0.21)	2.03(0.18)	0.93(0.17)	2.05(0.19)	0.93(0.18)	2.04(0.21)	0.94(0.20)
β_{18}	0	0.00(0.14)	0.96(0.14)	0.00(0.12)	0.96(0.13)	0.00(0.13)	0.95(0.13)	0.01(0.14)	0.96(0.14)
β_{19}	2	2.04(0.21)	0.94(0.21)	2.03(0.18)	0.93(0.17)	2.04(0.19)	0.94(0.18)	2.04(0.20)	0.96(0.20)
β_{110}	0	0.00(0.15)	0.94(0.14)	0.00(0.13)	0.96(0.13)	0.01(0.13)	0.95(0.13)	0.00(0.14)	0.95(0.14)
β_{20}	-6	-6.13(0.49)	0.95(0.48)	-6.10(0.36)	0.96(0.36)	-6.10(0.41)	0.95(0.39)	-6.12(0.43)	0.95(0.43)
β_{21}	-3	-3.07(0.30)	0.95(0.29)	-3.05(0.24)	0.94(0.24)	-3.05(0.26)	0.94(0.25)	-3.06(0.27)	0.94(0.27)
β_{22}	0	0.00(0.16)	0.95(0.16)	0.01(0.14)	0.95(0.14)	0.00(0.15)	0.95(0.15)	0.00(0.15)	0.95(0.15)
β_{23}	-3	-3.07(0.30)	0.96(0.29)	-3.05(0.24)	0.94(0.24)	-3.05(0.26)	0.94(0.25)	-3.06(0.27)	0.94(0.27)
β_{24}	0	0.00(0.16)	0.96(0.16)	0.00(0.13)	0.94(0.14)	0.00(0.14)	0.95(0.15)	0.01(0.15)	0.95(0.16)
β_{25}	2	2.04(0.21)	0.97(0.23)	2.04(0.19)	0.96(0.19)	2.04(0.20)	0.95(0.20)	2.04(0.20)	0.95(0.20)
β_{26}	0	0.01(0.17)	0.94(0.16)	0.01(0.15)	0.93(0.14)	0.01(0.15)	0.95(0.15)	0.00(0.16)	0.94(0.16)
β_{27}	3	3.06(0.29)	0.95(0.29)	3.04(0.24)	0.95(0.24)	3.04(0.26)	0.94(0.25)	3.05(0.27)	0.95(0.27)
β_{28}	0	0.01(0.16)	0.96(0.16)	0.00(0.14)	0.96(0.14)	0.00(0.15)	0.95(0.15)	0.01(0.15)	0.96(0.16)
β_{29}	0	0.00(0.16)	0.95(0.16)	-0.01(0.14)	0.94(0.14)	-0.01(0.15)	0.95(0.15)	-0.01(0.16)	0.96(0.15)
β_{210}	0	0.01(0.16)	0.95(0.16)	0.01(0.14)	0.94(0.14)	0.01(0.15)	0.95(0.15)	0.01(0.15)	0.94(0.16)
δ	0.3	0.29(0.09)	0.98(0.13)	0.28(0.07)	0.99(0.08)	0.29(0.07)	0.98(0.09)	0.29(0.07)	0.99(0.11)
$S_{e:1}$	0.95	–	–	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.94(0.01)	0.95(0.02)	0.94(0.01)
$S_{e:2}$	0.95	–	–	0.95(0.01)	0.95(0.01)	0.95(0.02)	0.93(0.01)	0.95(0.02)	0.91(0.02)
$S_{p:1}$	0.95	–	–	0.95(0.01)	0.96(0.01)	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.92(0.01)
$S_{p:2}$	0.95	–	–	0.95(0.01)	0.96(0.01)	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)

TABLE 4 The average prediction error PE×100 and the selection rate (SR) value (provided in parenthesis) of the maximum likelihood estimate and the shrinkage estimates under the tuning parameter criterion of Akaike information criterion (AIC), Bayesian information criterion (BIC), and extended regularized information criterion (ERIC) over 500 replications under S1;-S3 across individual testing (IT) and the State Hygienic Laboratory pooling with $c = 2$, 5, and 10. Recall that the SR is defined to be the proportion of the true model being exactly selected by a shrinkage estimator. The highest SR value under each setting is underlined

		IT	$c = 2$	$c = 5$	$c = 10$
Setting	Estimate	PE×100(SR)	PE×100(SR)	PE×100(SR)	PE×100(SR)
S1	MLE	0.148(0.000)	0.126(0.000)	0.130(0.000)	0.142(0.000)
	AIC	0.106(0.414)	0.092(0.430)	0.092(0.442)	0.102(0.462)
	BIC	0.079(0.910)	0.071(0.908)	0.073(0.926)	0.083(0.898)
	ERIC	0.085(0.724)	0.075(0.736)	0.076(0.744)	0.085(0.734)
S2	MLE	0.133(0.000)	0.106(0.000)	0.117(0.000)	0.121(0.000)
	AIC	0.095(0.414)	0.074(0.414)	0.084(0.436)	0.087(0.418)
	BIC	0.074(0.908)	0.059(0.910)	0.067(0.892)	0.069(0.876)
	ERIC	0.084(0.702)	0.064(0.696)	0.074(0.702)	0.074(0.702)
S3	MLE	0.284(0.000)	0.231(0.000)	0.250(0.000)	0.266(0.000)
	AIC	0.193(0.266)	0.160(0.294)	0.175(0.274)	0.184(0.298)
	BIC	0.158(0.818)	0.130(0.820)	0.145(0.786)	0.153(0.808)
	ERIC	0.183(0.428)	0.150(0.448)	0.163(0.420)	0.170(0.448)

compromising regression inference, the SHL pooling produces more precise inference. In addition, one must note that these advantages are achieved with a less amount of costs and a larger number of parameters to be estimated. This finding could be very encouraging to laboratories that are not using pooling (such as the NPHL).

5.2 | A CT/NG screening data set

To further encourage the use of pooling, we analyze a data set collected from the NPHL which currently uses individual testing for the CT/NG screening. We will illustrate, if switching from individual testing to the two-stage hierarchical pooling used by the SHL, what benefits could be achieved for regression. To do so, we first reiterate how the SHL is using the pooling protocol.[26] Only female swab specimens are screened using the SHL pooling protocol. The testing is carried out by the TECAN DTS platform with the Aptima Combo 2 assay. The platform is calibrated for a group size $c = 4$. The sensitivity and specificity of the assay are $S_{e:1} = 0.942$ ($S_{e:2} = 0.992$) and $S_{p:1} = 0.976$ ($S_{p:2} = 0.987$) for CT (NG), respectively (Gen-Probe Inc, San Diego, CA).

In 2009, 14 530 female swab specimens were tested individually in the NPHL. The employed assay was also the Aptima Combo 2 Assay. We are provided with the diagnosed results of each specimen for CT and NG. Based on these diagnoses, the approximated prevalence of CT and NG are 0.069 and 0.013, respectively. To reveal the benefits of pooling, we mimic the SHL screening practice in the most realistic way. We use a group size $c = 4$, which is used by the SHL. Then, we construct pools by assigning specimens according to their arrival time at the NPHL. Because the arrival time of specimens at the NPHL are random, our way of pooling is also random. We treat the diagnoses as "true" statuses and simulate a two-stage group testing data set using the above testing error rates. For comparison, we also simulate an individual testing data set using the same testing error rates. The considered covariates include age, prenatal, symptoms, cervical friability, pelvic inflammatory disease, cervicitis, multiple partners, and new partner in the last 90 days, and contact with someone who has an STD. All covariates, except age, are binary. With these covariates on each individual, we first fit the individual diagnoses results by viewing them as the truth. The resulting estimates are used as the "reference" estimates. We then fit the individual testing data and the two-stage group testing data using the regression and variable selection methods previously described. In our analysis, we standardize age and code dichotomous covariates as either −0.5 or 0.5.

Table 5 summarizes the parameter estimates and variable selection results. The estimates from both testing protocols are close to the "reference" estimates, but the SEs under $c = 4$ are uniformly less than the ones under individual testing. The testing error rates are estimated accurately from the group testing data. In terms of variable selection, the reference shrinkage estimates identified different sets of significant risk factors for the two infections, where prenatal is significant to CT but not to NG. The same results are identified by the three shrinkage estimates based on the group testing data. However, based on the individual testing data, none of the three shrinkage estimates can select prenatal for CT. These comparisons reinforce our conclusion that, in addition to a significant cost reduction (ie, it saves $14\,530 - 7737 = 6793$ tests), the two-stage pooling protocol leads to more precise inference than individual testing while estimating the testing error rates simultaneously. In addition, we have considered randomly assigning individuals into groups as in Section 5.1 and used group sizes varying from 2 to 10. The supplementary materials include these results, which reinforce the aforementioned conclusion on the advantages of the two-stage pooling protocol when compared with individual testing. We believe these numerical findings could encourage more laboratories to consider the two-stage pooling protocol.

6 | DISCUSSION

Motivated by the SHL CT/NG screening practice, we have developed a regression method for the two-stage hierarchical pooling data. Our proposed technique jointly models the unobserved individual disease statuses and produces interpretable marginal inference for each infection. The assay sensitivity and specificity for each infection can be estimated as well. In addition, we further developed a shrinkage estimator to consistently select truly relevant risk factors for each infection. To disseminate this work, code, written in R, that implements our new methodology, is available upon request.

From the simulation studies and the CT/NG screening data analysis, it is exciting to observe that, as compared with individual testing, the SHL pooling protocol can significantly reduce cost and yet produce more efficient regression estimators. An interesting future project would be to theoretically investigate how to construct groups to obtain the most efficient regression estimators for each infection within a budget limit. Intuitively, individuals with high probabilities of being infected should be tested individually and those with low probabilities could be tested in pools. However, what is the criterion to differentiate between high and low probabilities? How to know these probabilities before the screening? For those tested in pools, what is the optimal pool size that should be used for inference? These are interesting but challenging questions to be answered in future works. Possible guidance could be found in aforementioned studies.[17,46]

In our simulation studies, we used a Gumbel copula. We chose it for two reasons: (1) When compared to Gaussian copulas, it has an analytic expression that facilitates the computation, and (2) it is able to deliver robust estimates of

TABLE 5 The NPHL screening data analysis: parameter estimates (MLE), estimated standard errors (SE) and variable selection results (using the AIC, BIC, and ERIC criterion) from the reference estimates (Reference), individual testing estimates (IT), and the SHL pooling estimates with a group size 4 ($c = 4$). The number of tests under each is provided as well

Number of tests		**Reference**				**IT**				$c = 4$			
		–				**14 530**				**7737**			
		MLE(SE)	**AIC**	**BIC**	**ERIC**	**MLE(SE)**	**AIC**	**BIC**	**ERIC**	**MLE(SE)**	**AIC**	**BIC**	**ERIC**
CT	Intercept	-1.382(0.241)	–	–	–	-1.528(0.286)	–	–	–	-1.269(0.262)	–	–	–
	Age	-0.559(0.045)	✓	✓	✓	-0.535(0.057)	✓	✓	✓	-0.561(0.051)	✓	✓	✓
	Prenatal	0.390(0.220)	✓	✓	✓	0.141(0.291)	×	×	×	0.480(0.229)	✓	✓	✓
	Symptoms	0.356(0.079)	✓	✓	✓	0.324(0.095)	✓	✓	✓	0.356(0.088)	✓	✓	✓
	Cervical F	0.065(0.163)				-0.058(0.202)				0.003(0.182)			
	PID	0.443(0.392)	✓	✓	✓	0.443(0.448)	✓	✓	✓	0.492(0.427)	✓	✓	✓
	Cervicitis	0.611(0.106)	✓	✓	✓	0.746(0.118)	✓	✓	✓	0.645(0.116)	✓	✓	✓
	Multipartner	0.476(0.099)	✓	✓	✓	0.522(0.116)	✓	✓	✓	0.532(0.109)	✓	✓	✓
	New partner	-0.069(0.091)				-0.205(0.116)	✓			-0.067(0.102)			
	Contact STD	1.006(0.098)	✓	✓	✓	1.023(0.111)	✓	✓	✓	1.048(0.108)	✓	✓	✓
NG	Intercept	-2.426(0.416)	–	–	–	-2.727(0.595)	–	–	–	-2.683(0.507)	–	–	–
	Age	-0.251(0.083)	✓	✓	✓	-0.278(0.112)	✓	✓	×	-0.258(0.087)	✓	✓	✓
	Prenatal	0.283(0.591)				0.003(0.929)				-0.073(0.750)			
	Symptoms	1.202(0.164)	✓	✓	✓	1.176(0.219)	✓	✓	✓	1.234(0.174)	✓	✓	✓
	Cervical F	0.277(0.288)				0.290(0.327)	✓	✓	✓	0.270(0.301)			
	PID	1.032(0.496)	✓	✓	✓	0.719(0.635)	✓	✓	✓	0.879(0.554)	✓	✓	✓
	Cervicitis	0.625(0.199)	✓	✓	✓	0.746(0.225)	✓	✓	✓	0.712(0.201)	✓	✓	✓
	Multipartner	1.070(0.177)	✓	✓	✓	0.894(0.216)	✓	✓	✓	1.106(0.185)	✓	✓	✓
	New partner	-0.130(0.189)				-0.060(0.229)				-0.127(0.198)			
	Contact STD	1.405(0.173)	✓	✓	✓	1.208(0.216)	✓	✓	✓	1.402(0.180)	✓	✓	✓
	δ	0.573(0.030)	–	–	–	0.604(0.042)	–	–	–	0.563(0.033)	–	–	–
	$S_{e:1} = 0.942$	–	–	–	–	–	–	–	–	0.922(0.016)	–	–	–
	$S_{e:2} = 0.992$	–	–	–	–	–	–	–	–	0.989(0.029)	–	–	–
	$S_{p:1} = 0.976$	–	–	–	–	–	–	–	–	0.974(0.004)	–	–	–
	$S_{p:2} = 0.987$	–	–	–	–	–	–	–	–	0.985(0.002)	–	–	–

the regression coefficients and misclassification parameters even when the true copula is not Gumbel. To reveal this robustness, we have included a simulation study in the supplementary materials. In practice, users are welcome to choose other copulas, such as Gaussian, Clayton, or Frank.[33] Besides, the logistic function for g_k could also be changed to the inverse of the link in probit or complementary log-log models. Our GEM algorithm has the generality to incorporate those choices.

Although this work mainly focuses on two infections, the model can be extended to incorporate more infections. For example, suppose there are three infections. We have $\widetilde{\boldsymbol{Y}}_{ij} = (\widetilde{Y}_{ij1}, \widetilde{Y}_{ij2}, \widetilde{Y}_{ij3})^{\mathrm{T}}$. A joint model for $\widetilde{\boldsymbol{Y}}_{ij}|\boldsymbol{x}_{ij}$ is built by assuming that there exists a random vector $\boldsymbol{U}_{ij} = (U_{ij1}, U_{ij2}, U_{ij3})^{\mathrm{T}}$, of which the distribution function is a three-dimensional copula $C(u_1, u_2, u_3|\delta)$, such that the event $\{\widetilde{Y}_{ijk} = 1|\boldsymbol{x}_{ij}\}$ is equivalent to $\{U_{ijk} \leq g_k(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_k)\}$ for $k = 1, 2, 3$. Consequently, the marginal regression model (1) naturally holds for each disease, and the cell probabilities of $\widetilde{\boldsymbol{Y}}_{ij}|\boldsymbol{x}_{ij}$ can be calculated in terms of C, eg, $\mathrm{pr}(\widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 1, \widetilde{Y}_{ij3} = 0|\boldsymbol{x}_{ij}) = C\{g_1(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1), g_2(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2), 1|\delta\} - C\{g_1(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1), g_2(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2), g_3(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_3)|\delta\}$. Our GEM algorithm can be generalized to incorporate more than two infections as well. We omit details but include some simulation results in the supplementary materials to demonstrate this generalizability.

Lastly, we discuss the three assumptions (Assumptions 1–3) on the assay sensitivity and specificity and possible ways to relax them. For Assumption 1, when the assay utilizes the concentration level of a specific biological marker (biomarker) to make a diagnosis, mixing a positive specimen with negative ones could dilute the concentration level and affect the assay sensitivity and specificity significantly when group size changes. This "dilution effect" can be taken into consideration if the distribution of the biomarker concentration is provided in advance.[47,48] To relax Assumption 2, one could use a multinomial distribution to account for the cross-disease dependency of the testing outcomes when the true statuses are given. Then, the number of misclassification parameters increases from 4 to 12 when the number of diseases is two. One could modify the GEM algorithm to estimate the 12 parameters along with the regression. However, some of these parameters may require an impractical large sample size to be accurately estimated. The last assumption can be relaxed by assuming a covariate-adjusted model for misclassification parameters.[49] However, caution must be taken for model identifiability when the covariate-adjusted misclassification parameters are to be estimated along with the regression.

ACKNOWLEDGEMENTS

This article was supported by the National Institutes of Health under grant R03 AI135614. The authors thank an Associate Editor and three reviewers for their constructive comments that lead to a better presentation of this work, and Drs. Joshua M. Tebbs and Christopher S. McMahan for their insightful comments. The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Dewei Wang https://orcid.org/0000-0003-0822-8563

REFERENCES

1. Centers for Disease Control and Prevention. 2016 STD surveillance report. https://www.cdc.gov/std/stats16/default.htm. Accessed April 2018.
2. Lewis JL, Lockary VM, Kobic S. Cost savings and increased efficiency using a stratified specimen pooling strategy for Chlamydia trachomatis and Neisseria gonorrhoeae. *Sex Transm Dis*. 2012;39(1):46-48.
3. Samoff E, Koumans EH, Markowitz LE, et al. Association of Chlamydia trachomatis with persistence of high-risk types of human papillomavirus in a cohort of female adolescents. *Am J Epidemiol*. 2005;162(7):668-675.
4. Centers for Disease Control and Prevention. STDs & infertility. https://www.cdc.gov/std/infertility/default.htm. Accessed April 2018.
5. Jirsa S. Pooling specimens: a decade of successful cost savings. In: Proceedings of the National STD Prevention Conference; 2008; Chicago, IL.
6. Dorfman R. The detection of defective members of large populations. *Ann Math Stat*. 1943;14(4):436-440.
7. Stramer SL, Krysztof DE, Brodsky JP, et al. Comparative analysis of triplex nucleic acid test assays in united states blood donors. *Transfusion*. 2013;53:2525-2537.

8. Edouard S, Prudent E, Gautret P, Memish ZA, Raoult D. Cost-effective pooling of DNA from nasopharyngeal swab samples for large-scale detection of bacteria by real-time PCR. *J Clin Microbiol*. 2015;53(3):1002-1004.
9. Hill JA, HallSedlak R, Magaret A, et al. Efficient identification of inherited chromosomally integrated human herpesvirus 6 using specimen pooling. *J Clin Virol*. 2016;77:71-76.
10. Gastwirth JL. The efficiency of pooling in the detection of rare mutations. *Am J Hum Genet*. 2000;67(4):1036-1039.
11. Muñoz-Zanzi CA, Johnson WO, Thurmond MC, Hietala SK. Pooled-sample testing as a herd-screening tool for detection of bovine viral diarrhea virus persistently infected cattle. *J Vet Diagn Investig*. 2000;12(3):195-203.
12. Venette RC, Moon RD, Hutchison WD. Strategies and statistics of sampling for rare individuals. *Annu Rev Entomol*. 2002;47(1):143-174.
13. Dodd RY, Notari EP IV, Stramer SL. Current prevalence and incidence of infectious disease markers and estimated window-period risk in the American Red Cross donor population. *Transfusion*. 2002;42(8):975-979.
14. Remlinger KS, Hughes-Oliver JM, Young SS, Lam RL. Statistical design of pools using optimal coverage and minimal collision. *Technometrics*. 2006;48(1):133-143.
15. Kim H-Y, Hudgens MG, Dreyfuss JM, Westreich DJ, Pilcher CD. Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*. 2007;63(4):1152-1163.
16. Liu A, Liu C, Zhang Z, Albert PS. Optimality of group testing in the presence of misclassification. *Biometrika*. 2012;99(1):245-251.
17. Huang S-H, Huang M-NL, Shedden K, Wong WK. Optimal group testing designs for estimating prevalence with uncertain testing errors. *J Royal Stat Soc Ser B*. 2017;79(5):1547-1563.
18. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000;56(4):1126-1133.
19. Xie M. Regression analysis of group testing samples. *Statist Med*. 2001;20(13):1957-1969.
20. Chen P, Tebbs JM, Bilder CR. Group testing regression models with fixed and random effects. *Biometrics*. 2009;65(4):1270-1278.
21. McMahan CS, Tebbs JM, Hanson TE, Bilder CR. Bayesian regression for group testing data. *Biometrics*. 2017;73:1443-1452.
22. Delaigle A, Meister A. Nonparametric regression analysis for group testing data. *J Am Stat Assoc*. 2011;106(494):640-650.
23. Delaigle A, Hall P, Wishart JR. New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data. *Biometrika*. 2014;101(3):567-585.
24. Xiao X, Zhai J, Zeng J, Tian C, Wu H, Yu Y. Comparative evaluation of a triplex nucleic acid test for detection of HBV DNA, HCV RNA, and HIV-1 RNA, with the Procleix Tigris System. *J Virol Methods*. 2013;187(2):357-361.
25. Hughes-Oliver JM, Rosenberger WF. Efficient estimation of the prevalence of multiple rare traits. *Biometrika*. 2000;87(2):315-327.
26. Tebbs JM, McMahan CS, Bilder CR. Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project. *Biometrics*. 2013;69(4):1064-1073.
27. Warasi MS, Tebbs JM, McMahan CS, Bilder CR. Estimating the prevalence of multiple diseases from two-stage hierarchical pooling. *Statist Med*. 2016;35(21):3851-3864.
28. Li Q, Liu A, Xiong W. D-optimality of group testing for joint estimation of correlated rate diseases with misclassification. *Statistica Sinica*. 2017;27(2):823-838.
29. Zhang B, Bilder CR, Tebbs JM. Regression analysis for multiple-disease group testing data. *Statist Med*. 2013;32(28):4954-4966.
30. Wu CFJ. On the convergence properties of the em algorithm. *Ann Stat*. 1983;11:95-103.
31. Neal RM, Hinton GE. A view of the em algorithm that justifies incremental, sparse, and other variants. In: *Learning in Graphical Models*. Dordrecht, The Netherlands: Springer Science & Business Media; 1998:355-368.
32. Gregory KB, Wang D, McMahan CS. Adaptive elastic net for group testing. *Biometrics*. 2019;75:13-23.
33. Nelsen RB. *An Introduction to Copulas*. New York, NY: Springer Science & Business Media; 2007.
34. Lehmann EL. *Theory of Point Estimation*: Pacific Grove, CA: Wadsworth and Brooks-Cole; 1983.
35. Louis TA. Finding the observed information matrix when using the EM algorithm. *J Royal Stat Soc Ser B Methodol*. 1982;44:226-233.
36. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429.
37. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407-499.
38. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360.
39. Wang H, Leng C. Unified lasso estimation by least squares approximation. *J Am Stat Assoc*. 2007;102(479):1039-1048.
40. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
41. Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. *J Royal Stat Soc Ser B*. 2009;71(3):671-683.
42. Minnier J, Tian L, Cai T. A perturbation method for inference on regularized regression estimates. *J Am Stat Assoc*. 2011;106(496):1371-1382.
43. Gumbel EJ. Bivariate exponential distributions. *J Am Stat Assoc*. 1960;55(292):698-707.
44. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control*. 1974;19(6):716-723.
45. Hui FKC, Warton DI, Foster SD. Tuning parameter selection for the adaptive lasso using ERIC. *J Am Stat Assoc*. 2015;110(509):262-269.
46. McMahan CS, Tebbs JM, Bilder CR. Informative Dorfman screening. *Biometrics*. 2012;68(1):287-296.
47. Wang D, McMahan CS, Gallagher CM. A general parametric regression framework for group testing data with dilution effects. *Statist Med*. 2015;34(27):3606-3621.
48. Wang D, McMahan CS, Tebbs JM, Bilder CR. Group testing case identification with biomarker information. *Comput Stat Data Anal*. 2018;122:156-166.
49. Janes H, Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *Am J Epidemiol*. 2008;168(1):89-97.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary materials are available along with the submission. These materials contain a numerical comparison showing that ignoring the retesting outcomes could inflate the variance of estimators of the regression coefficients in Section 2, the observed log-likelihood function, a numerical study showing the computational advantages of the GEM algorithm, detailed derivations of the E-step and the observed data information matrix introduced in Section 3, additional numerical results for other values of $S_{e:k}$ and $S_{p:k}$ (Section 5.1), extensions of our method to fit individual testing data as discussed in Section 5.1, additional results of the real data analysis in Section 5.2, and simulation studies that reveal the robustness of the Gumbel copula and demonstrate the generalizability of our method to more than two infections in Section 6.

How to cite this article: Lin J, Wang D, Zheng Q. Regression analysis and variable selection for two-stage multiple-infection group testing data. *Statistics in Medicine*. 2019;1–15. https://doi.org/10.1002/sim.8311